Hot Mix Asphalt Volumetric Optimization using Artificial Intelligence

by

Jesus USECHE

THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE IN PARTIAL FULFILLMENT FOR A MASTER'S DEGREE WITH THESIS IN CONSTRUCTION ENGINEERING
M.A. Sc.

MONTREAL, NOVEMBER 24TH, 2022

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. Alan Carter , Thesis Supervisor
Department of Construction engineering, École de technologie supérieure

Mr. Freddy Sanchez-Leal Thesis Co-supervisor
Ramcodes company

Mr. Ulrich Aïvodji, President of the Board of Examiners
Department of Software Information Technology Engineering, École de technologie supérieure

Mr. Éric Lachance-Tremblay, Member of the jury
Department of Construction engineering, École de technologie supérieure

THIS THESIS WAS PRENSENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND PUBLIC

MONTREAL, OCTOBER 19TH, 2022

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGMENT

I would like to thank my family: my beloved wife Ingrid Segura and my father Juan de Jesus Useche for supporting me psychologically and spiritually throughout writing this thesis and my life in general.

Throughout the writing of this work, I have received a great deal of support and assistance.

I would like to show my greatest appreciation to my advisor Prof. Alan Carter. I can't say thank you enough for his tremendous patience and support by having empathy, understanding my situation. His trust in me, gave me the motivation necessary to persist despite the constrains. Without his encouragement and guidance, this thesis would not have materialized. Moreover, I would like to thank Gabriel Orozco for his appropriate technical help. I would also like to thank my tutor, my always professor and my dear friend, Freddy Sanchez-Leal, for his valuable guidance throughout my studies and professional life. You provided me with the research tools (critical thinking) that I needed to choose the right direction and successfully complete this study.

Finally, I would like to acknowledge my colleague Juan David Silva whose expertise was an invaluable resource to apply machine learning concepts into the field of Hot Mix Asphalt volumetric design.

Hot mix asphalt volumetric optimization using artificial intelligence

Jesus Useche

RÉSUMÉ

La conception volumétrique des enrobés à chaud est une procédure empirique qui demande du temps et des efforts. Cette procédure est une pratique habituelle lorsque de possibles changements dans les agrégats et les propriétés du bitume imposent aux usines d'enrobage l'ajustement de la formule de travail. Cette conception volumétrique est basée sur des spécifications dans lesquelles la granulométrie est limitée, ne peut pas fournir un lien clair entre l'état des vides souhaités et le mélange des agrégats. Donc, un matériau d'enrobé est produit avec une performance incertaine auquel un test de validation est essentiel.

Ce travail compare 4 facteurs granulométriques dans lesquels la granulométrie de l'enrobé est quantifiée sous forme de nombre. Cette évaluation permet de savoir quel facteur granulométrique a la corrélation la plus élevée concernant les propriétés volumétriques de l'enrobé, en réduisant ainsi les efforts impliqués dans la conception empirique en raison d'un simple calcul. Ces facteurs pourraient produire une estimation efficace l'état des vides de l'enrobe.

Cette étude applique des réseaux de neurones profonds (deep neural networks : DNN) basés sur une méthode d'apprentissage automatique supervisé pour prédire deux propriétés volumétriques : 1) la densité maximale de l'enrobé (Dmm) laquelle est déterminée avec l'essai Rice et 2) la densité brute de l'enrobé (Dmb) dans laquelle le paquet d'agrégats est compacté avec la presse à cisaillement giratoire sous le standard superpave et la teneur en bitume produit des changements de densité de l'enrobé au même niveau d'énergie de compactage à la même granulométrie. Ces réseaux de neurones sont construits sur la base de propriétés d'agrégat et du bitume comme variables indépendantes et de propriétés volumétriques de l'enrobé en tant que variables dépendantes. En plus, ces réseaux de neurones sont développés en utilisant des données collectées de la base de données de Long-Term Pavement Performance (LTPP). Cette recherche explore établir avec précision des liens entre les propriétés des agrégats et du bitume avec les propriétés volumétriques des enrobés.

Une corrélation robuste a été trouvée entre la propriété volumétrique du combiné granulométrique (Dgb) et la prédiction de la Dmm. De plus, la prédiction de la Dmm basée sous une généralisation correctement évaluée du phénomène à l'aide de réseaux de neurones, pourrait permettre aux concepteurs d'estimer avec précision la propriété volumétrique (Dmm), en évitant ainsi de réaliser ce test dans la méthode de conception d'enrobés pour calculer les valeurs volumétriques.

Mots-clés: Enrobés; Conception volumétrique; Optimisation; intelligence artificielle; Réseaux de neurones profonds pour la conception des enrobés.

Hot mix asphalt volumetric optimization using artificial intelligence

Jesus USECHE

ABSTRACT

The empirical procedure of HMA volumetric design requires time and effort. This procedure is a common practice when adjusting the job formula due to eventual changes in aggregates and binder properties imposing upon asphalt plants. This empirical method is based on gradation specifications limits in which correlations between the desire state of voids and aggregate gradation remain unclear, thereby producing a material (HMA) with uncertain performance at which validation test is essential.

This work compares 4 gradation factors in which aggregate gradation is quantified as a number. This evaluation permits knowing which gradation factor has the highest correlation with HMA volumetric properties, thereby reducing the endeavor implied in the empirical design due to a simple calculation of the factor could produce an effective estimation of volumetric properties.

This study applies deep neural networks (DNNs) based on supervised machine learning method to predict two volumetric properties: 1) the maximum specific gravity (Gmm) which is obtained by running the Rice test and 2) the bulk specific gravity (Gmb) in which the aggregate packing is compacted when SGC test is applied, and asphalt binder content produces bulk density changes at the same level of energy compaction for the same aggregate blend. These neural networks are built based on aggregate and binder properties as input variables and HMA volumetric properties as output variables. These DNNs are training to use raw data collected from the Long-Term Pavement Performance (LTPP) database. This research explores how to accurately establish links between aggregate and binder properties with HMA volumetric properties (Gmb and Gmm).

A strong correlation was found between the aggregate volumetric property Gsb and Gmm prediction. Moreover, the generalization of the Rice test phenomenon to adequately predict Gmm using DNNs could allow designers to precisely estimate the HMA volumetric property (Gmm), thus avoiding running this test to calculate the volumetric values.

Keywords: Asphalt mix; Artificial intelligence in HMA; Deep neural network for mix asphalt design; Hot mix Asphalt volumetric optimization

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Page

LIST OF ABREVIATIONS

AI	Artificial Intelligent
ANN	Artificial Neural Network
BMD	Balanced Mix Design
DASR	Dominant Aggregate Size Range
Dmax	Maximum diameter
DNNs	Deep Neural Networks
Fp	The Characteristic Factor
GI	Gradation index
Gmb	Bulk Specific Gravity
Gmm	Maximum Specific Gravity
G/S	The Gravel-to-Sand ratio
Gsb	Bulk specific gravity for the total aggregate
Gse	Effective specific gravity of the total $aggregate
HMA	Hot Mix Asphalt
IC	Interstitial Components
IV	Interstitial Voids
LTPP	Long-Term Pavement Performance
Lr	Learning rate
MAE	Mean Absolute Error
ML	Machine Learning
MLTs	Machine Learning Techniques
NMPS	Nominal Maximum Particle Size
PCS	Primary Control Sieve
RAMCODES	Rational Methodology for Compacted Geomaterial´s Densification and Strength analysis
RMSE	Root Mean Square Error
RAP	Reclaim Asphalt Pavement
RT	Rice Test
SGC	Superpave Gyratory Compactor

SCS	Secondary Control Sieve
TCS	Tertiary Control Sieve
TG	Total Granulometry

INTRODUCTION

Understanding the extent to which aggregates are arranged into compacted Hot Mix Asphalt (HMA) in order to possibly achieve an appropriate performance has attracted the attention of several researchers to optimize the HMA design. In effect, volumetric approaches have been considered based on the analysis of aggregate blend porosity to explain in what manner fine particles arrange into the space generated by the coarse particles. This arrangement has been measured evaluating how voids and total volume of bulk change when the proportion of fine-coarse particles changes into the aggregate blend.

Moreover, this packing phenomenon has also been evaluated as a theoretical two-dimension method in which interstitial voids generated by coarse particles can be measured in terms of the diameter of the coarse particles referred to the Bailey method. Another method has been the evaluation the porosity of the dominant aggregate size range (DASR) with interstitial components (IC). However, the main pattern of all these studies is the assumption that the interaction of aggregates – both coarse and fine ones – into the HMA is in a dry condition, thereby avoiding the real action of the asphalt binder content into the HMA, which at the same aggregate blend, produces bulk density changes at equal levels of energy compaction. This evidence shows an important limitation wherein the merely extrapolating the volumetric results from unbound to bound materials, remaining far from the reality of HMA packing phenomenon.

One approach to adequately address the HMA packing evaluation to optimize the HMA volumetric design is developing models under which the asphalt binder helps coarse particles to arrange by lubrication effect reducing interstitial voids (IV) while increasing bulk densification at the same compaction energy just until the binder begins to disrupt the structure aggregate or contact-contact particles. Nevertheless, limitations have been found when several independent variables should be considered by using traditional statistical method. It seems to be one of the reasons why researchers simplified, in a dry condition, the aggregate packing phenomenon, despite the evidence of bulk density changes at equal levels of energy compaction when asphalt binder changes.

Therefore, today the fast development of computers has allowed researchers to apply the Artificial Intelligent (AI) approach into several disciplines, including computer vision, speech and audio processing, natural language processing, robotics, bioinformatics and chemistry, video games, online advertising, finance, medicine and engineering to overcome limitations regarding: 1) traditional statistical methods and 2) empirical test. The former, traditional statistical methods, find difficulties correlating more than three variables regarding the phenomenon studied. The later empirical test limits the data collection due to the time effort toward running tests. Additionally, these empirical results could produce wrong conclusions because the present risk of the conditions test assumed differs from the real conditions, thus revealing the existence of confounding variables that interfere in the phenomenon studied.

The present research work attempts to adequately understand the aggregate packing arrangement of HMA opening the possibility to optimize the volumetric HMA design by using the Deep Neural Network (DNN) method if a sensitive analysis of input variables of the model can be performed. Thus, allowing observing which material (aggregates and asphalt binder) parameters have more influence (weight) into the Superpave gyratory compactor (SGC) to obtain the bulk specific gravity (Gmb) and the Rice test (RT) to obtain the maximum specific gravity (Gmm). Thus, a new evaluative approach is developed, in which HMA volumetric design could be optimized avoiding the traditional trial-and-error procedures. This supervised machine learning algorithm (DNN) can successfully generalize the phenomenon when large amounts of data can support it. This developed DNN is training using the database of The Federal Highway Administration Long-Term Pavement Performance (LTPP) program in which a large amount of research quality pavement performance information is collected and most of the standard tests like aggregate properties and HMA volumetric parameters are registered for the most important pavements of the United States.

The main objective of this work is to optimize the volumetric HMA design by predicting the two principal volumetric properties: 1) the Gmb under SGC test and 2) the Gmm under Rice test. Two models were developed by applying machine learning (ML) approach in which deep neural networks (DNN) were built considering the aggregate and binder properties as independent variables and the maximum specific gravity (Gmm) and bulk specific gravity of

the HMA (Gmb) as dependent variables. This DNN method permits observing which material properties have more influence into the HMA volumetric properties (Gmb and Gmm).
Thus, achieving an important contribution in which the two models developed are capable of handling aggregate and binder information to predict the (Gmm) and the (Gmb). This study is the one of the important step towards the goal of using AI in order to reduce the lab work and make the process more efficient by virtually automating crucial parts of the process in the traditional trial error HMA design.

In addition, some minor contributions are listed as follows.

1) Provide experimental support regarding the Gmm phenomenon:
   a) which gradation factors are the most influential;
   b) which material property are the most influential;
   c) which binder property are the most influential.
2) A pre-trained DNN model that can be used for future research on HMA performance prediction.

This document is organized in three chapters as follows:
In the first chapter, this research presents the state-of-the-art of the aggregates packing models in which HMA volumetric properties are evaluated. Next, a review of the common Machine Learning algorithms like Artificial Neural Networks are shown. Later, there is an evaluation of how some works where Machine Learning algorithms were applied in the field of HMA.
In the second chapter, the main methodology is described, which includes the data pairs set building, the secondary input data preprocessing, input variable selection criteria, selection model process and metric error. Finally, a validation of the optimum models and a sensitive (correlation) analyze were performed.
Experimental results are shown in the third chapter under which the proper metrics are presented for measuring the performance of the two models developed. In addition, results of the correlation analyze are presented, thus revealing which inputs have highest influence into the phenomenon.

A summary of findings and insights is discussed in this chapter, thereby linking the results with the theories mentioned in the first chapter regarding strengths and weaknesses of this research. Finally, conclusions are drawn with some hints for future works.

CHAPTER 1

LITERATURE REVIEW

This chapter summarizes and describes the main theoretical concepts regarding aggregate packing, HMA design, machine learning techniques. It can be separated in two main parts, the first one composed by sections 1.1 to 1.3, describes how researchers have been studied the aggregate packing and how HMAs have been designed. The second part starts in section 1.4 and describes the machine learning techniques and how these concepts have been applied in the field of HMA to address performance prediction.

1.1 Aggregate Packing

The aggregate packing in the field of HMA is the skeleton formed by different aggregate particles which the smallest parts find places into the interstitial voids (IV) generated by the biggest particles. In effect, this phenomenon of aggregate packing has attracted the interest of several researchers by which volumetric approaches have been studied. These volumetric approaches are based on the analysis of aggregate blend porosity to explain the way in which fine particles can be placed into the space generated by coarse particles. This aggregate arrangement has been measured by different ways. For example, some approaches measure this aggregate arrangement by evaluating directly how voids and total volume of bulk change when the proportion of fine-coarse particles changes into the aggregate blend (Perraton, Meunier, & Carter, 2007) (Shen & Yu, 2011).

Moreover, this aggregate packing phenomenon has been studied in other two approaches. The former, a theoretical two-dimension method in which interstitial voids generated by coarse can be measured in terms of the diameter of coarse particles (William R. Vavrik, Pine, & Carpenter, 2002) referred to the Bailey method; and the latter, the porosity evaluation of the dominant aggregate size range (DASR) and interstitial components (IC) (Kim, 2006), (Guarin, 2009), (Isola, 2014) . However, all these studies assume the packing phenomenon of HMA, only under the interaction of aggregates (coarse-fine) avoiding the action of the asphalt binder as an essential component. In effect, it is known that this asphalt binder content produces bulk

density changes at the same level of energy compaction for the same aggregate blend. This behaviour seems to show that merely extrapolating the volumetric results from unbound to bound materials remains far from the reality of HMA packing phenomenon.

Furthermore, compaction studies have shown robust difference between unbound and bound (HMA) materials. The former (Henderson, Herrington, Patrick, Kathirgamanathan, & Cook, 2011) provides strong evidence that particle orientation presents similar pattern effect regarding the average angle of inclination to the horizontal and vector magnitude in unbound materials, by using several compaction techniques like Marshal hammer, Gyratory compactor, Vibratory hammer, Hand operated vibrating roller, and Mikasa vibrating plate compactor. The latter (Tashman, Masad, Peterson, & Saleh, 2000) compared different conditions for bound (HMA) materials under which solely changing the angle of the gyratory compactor produces important changes in the vector magnitude of aggregate orientation. In fact, this vector magnitude influences the performance of HMA especially the stiffness (Masad, Tashman, Somedavan, & Little, 2002).

In summary, these studies clearly show that asphalt binder plays an important role into the aggregate packing of HMA. The former study seems to indicate that compaction methods did not affect unbound materials in contrast to the latter which showed that bound materials (HMA) are influenced by changing in compaction method.

One approach to adequately address the HMA packing evaluation is considering the hypothesis under which the asphalt binder helps coarse particles to arrange by lubrication effect, thereby reducing interstitial voids (IV) while increasing bulk densification at the same energy compaction. However, this densification behaviour is achieved just until the point in which the binder begins disrupting the structure aggregate or contact-contact particles of the coarse part, thus decreasing densification of the HMA bulk. Thus, a method is proposed in chapter 2 to overcome this limitation found in HMA optimization based on aggregate packing under dry condition assumption.

1.1.1 Bailey Method

The Bailey method is one of the most common optimization approaches of the aggregate packing in HMA volumetric design presented by (William Robert Vavrik, 2000) in which it is affirmed that the particle diameter ratio for the sized materials of 0.22 would be an adequate value for evaluating aggregate gradations. This method was developed by Bob Bailey of the Illinois Department of Transportation in the 80s (Aurilio, Pine, & Lum, 2005) which consists in controlling the mix volumetric properties to indirectly achieve performance (HMA compaction during construction, workability and segregation. In effect, Bailey's method establishes four (4) main principles:

1) defining properly which parts are the Coarse particles and which ones are the Fine particles,
2) evaluating the coarse fraction of the overall aggregate packing and the distribution of particle sizes based on the concept of large particles create voids,
3) evaluating the coarse part of the fine fraction regarding the overall fine fraction,
4) analyzing the fine part of the fine fraction regarding how this portion performs into the aggregate packing based on smaller particles fill voids created by coarse aggregates.

Effectively, the work of (William Robert Vavrik, 2000) implements control sieves into HMA design regarding The NMPS which is referred to one sieve larger than the first sieve that retains more than 10% in Superpave approach. This control sieve allows designers to find ratios as shown in Figure 1.1, with the object to analyze the performance of HMA which are:

1) Primary Control Sieve (PCS): it is identified as the division between the coarse and fine aggregate the sieve which properly define the particle sizes considered as coarse aggregate. This division is determined by multiplying the Nominal Maximum Particle Size (NMPS) by 0.22. Basically, there is a PCS according to the NMPS of the aggregate blend for each packing. Bailey's method concludes that four percent (4 %) change in PCS would be equivalent to one percent (1 %) changes in VMA. This gradation factor PCS is calculated by:
   a) $PCS = NMPS \times 0.22$
2) Secondary Control Sieve (SCS) separating the fine fraction of a combined gradation.

- SCS = PCS × 0.22

3) Tertiary Control Sieve (TCS): for the very fine particles filling the voids.
    - TCS = SCS × 0.22

By using these sieve control concepts, the method calculates some ratios to evaluate HMA volumetric properties:

1) Coarse Aggregate Ratio (CA): it is the coarse portion of the aggregate gradation, to evaluate the void structure in the coarse aggregate. This ratio is calculated based on the difference between The Half Sieve (HS) which is defined as half of the NMPS (NMPS × 0.5) and The PCS divided by 100-HS. This CA ratio permits elucidate that the higher the increment of CA ratio, the higher the VMA;
2) Coarse Portion of Fine Aggregate (FAc): it is the coarser portion interlock of the fine aggregate which also creates voids among particles, and the finer particles will fill these voids. FAc describes the interlock of the coarse portion in the fine aggregate. A change of 0.05 of this value would be equal to 1 % change in VMA or total voids. Basically, the higher the ratio the smaller the VMA. The FAc is calculated dividing the percentage (%) passing SCS by the one (%) passing PCS;
3) Fine Portion of Fine Aggregate (FAf): it is the finest material of the aggregate packing (including filler). The higher the Faf the smaller the VMA and a change of 0.05 of the value would be equal to one percent (1%) in VMA or total voids.

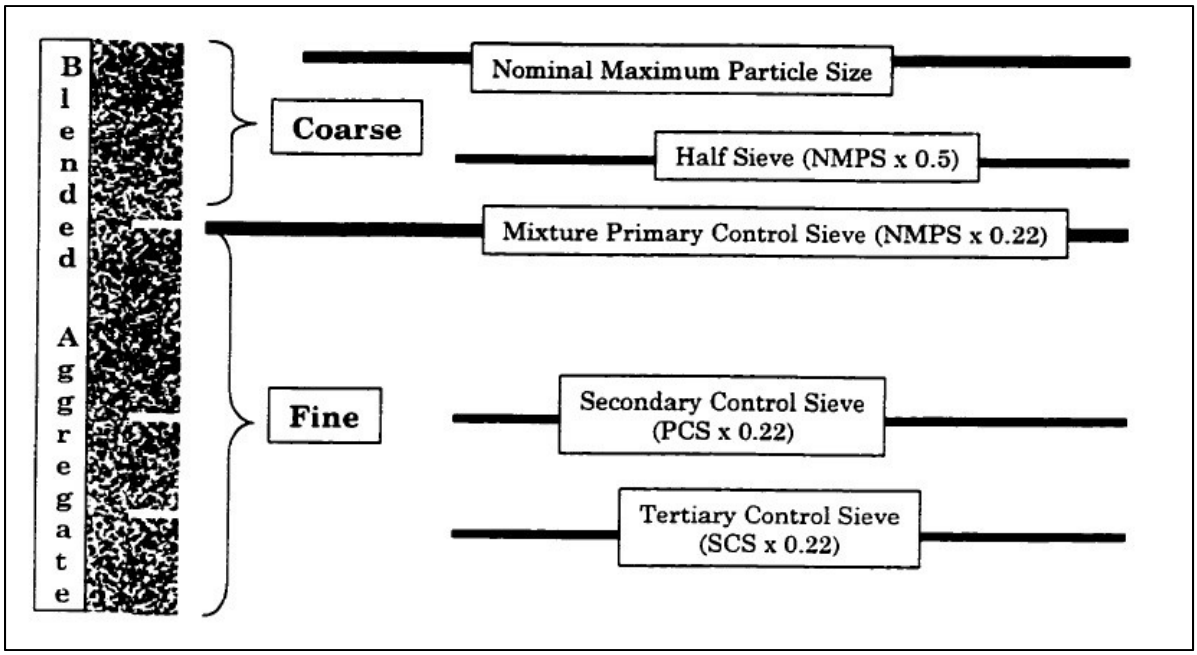


Figure 1. 1 Schematic Diagram of Control Sieves taken from William Robert Vavrik (2000)

1.1.2 Dominant Aggregate Size Range (DASR)

DASR is a theoretical aggregate packing concept developed by (Kim, 2006) which elucidates how the primary structural network of aggregates (coarse) performs based on gradations. This work shows that an interactive network could be achieved if the relative proportion larger/smaller of two contiguous size particles is within 70% from one sieve size and greater than 30% from the contiguous sieve. Moreover, this concept denotes that possible disruption interaction of coarse particles, produced by one particle size could be avoided if the relative proportion 70/30 is limited. This proportion concept of the aggregate packing indicates that the higher are the proportions large/smaller of 70/30, the more is the spacing of the smaller proportion, thereby swimming in the matrix with no interaction. However, if proportions are less than 70/30 (e.g., 60/40, 50/50, 40/60) each particle size retains a relatively constant spacing, thus being both part of the DASR due to interaction between particles.

In addition, all these contiguous particles which interact into the matrix are considered to act as a unit for the determination of porosity. This porosity of the mixture was the criterion for aggregate interlock. This criterion was determined to apply the soil mechanics concept in

which granular material porosity in a loose state could be constant between 45% and 50% despite the gradation distribution and particle size (Al-Mosawe & Thom, 2016). Furthermore, the study concluded that porosity less than 50% would produce particles contact with each other, thus achieving a mixture that performs well regarding resistance to deformation and cracking.

1.1.3 Discussion of Bailey Aggregate Packing Concepts

The Bailey, the method based on restrictions of the VMA regarding gradations changes, has been used as an alternative to optimize HMA design. However, clear limitations were found by (Al-Mosawe & Thom, 2016) when the Bailey factors were not efficient to define the differences between the gradations of each mixture studied. Thus, developing three more ratios or new Bailey factor had to be applied to overcome this constrains. In addition, the research of (Shen & Yu, 2011) elucidates that the manipulation of aggregate gradations in Bailey's method to reach the desire VMA still depends on an empirical trial mix design experiment to achieve initial VMA. Therefore, no direct correlation exists between the aggregate gradation and the expected VMA values. Thus, based on these limitations this research avoided apply bailey concept ratios as inputs into the two models developed and described in chapter 3.
Although the DASR concept seems to be a good approach regarding to the aggregate packing evaluation, this research avoided using the concept because one of the main objectives is the evaluation of gradation influence into the HMA volumetric properties.

1.2 Gradation Quantification Factors

Researchers have studied the way to establish a link between performance and aggregate gradation. For instance (F. Sánchez-Leal, 2003; Sanchez-Leal Jucá, De Campos, & Marinho, 2002) have developed a characterization factor Fp to characterize at first soils; however, this characterization concept was applied as an alternative for HMA quality control in plants (F. Sánchez-Leal, 2002; Sanchez Leal, Garnica Anguas, Gómez López, & Pérez García, 2002). Moreover, the study of (F. J. Sánchez-Leal, 2007) showed a robust correlation between the

ratio gravel-sand (G/S) and HMA performance. Furthermore, the study of (A. Setiawan, L. Suparma, & A. Mulyono, 2016a; A. Setiawan, L. B. Suparma, & A. T. Mulyono, 2016b; Setiawan, Suparma, & Mulyono, 2017) propose the GI gradation factor to correlate gradations with performance like tensile strength workability, and Marshall stability. Finally, the province of Quebec uses the TG for quality control in HMA production.

1.2.1 Rational Methodology for Compacted Geomaterial's Densification and Strength Analysis (RAMCODES)

Ramcodes method is a geotechniqueal approach proposed by (Sanchez Leal, Garnica Anguas, Gómez López, & Pérez García, 2002) under which the method considers the application of geotechniqueal concepts in order to analyze a compacted geomaterial. This approach allows engineers elucidating geomaterial combination in similar terms. The analysis method implies that the soils or rock concepts can be used for hydraulic cement concrete and HMA. The method is based on three postulated (F. J. Sánchez-Leal, 2007):

1) The term compacted geomaterial is referred to any combination of soil, soil cement, asphalt mix, or hydraulic cement mix (cement soil), thus applying concepts and criteria developed for each material to the rest.
2) The aggregate gradation and the specific surface of fines should be considered into the classification system for geomaterials as the main inherent factors that influence into mechanical and hydraulic behaviour.
3) The classification system should be quantitative (operationalization as a number in a continuous scale), thereby correlating response parameters with classification could be achieved; opposite to the traditional qualitative classification systems used in which no link can be adequately developed between classification and performance.

Above-explained RAMCODES principles can be readily applied to link gradations to behaviour of Unbound Granular Materials (UGM) with low fines, and to HMA. First, UGM or hot mix asphalt gradation is "quantified" by turning gradation data into a couple of parameters, namely, shape factor (n), and maximum diameter (Dmax), by the application of a simple allometric model. This method is applicable to UGM and HMA with continuous

gradations. RAMCODES uses shape factor (n) vs. maximum diameter (Dmax) chart, instead of the traditional particle diameter vs. % passing chart (gradation curve). By this, a single point is obtained representing the gradation curve, while a gradation specification band is transformed into a rectangle or box as seen in Figure 1.2.

This n-Dmax chart that displays gradations and specification bands is referred to "gradation chart" in RAMCODES. This gradation chart can be used to represent several HMA gradations and specifications, and associate to them mechanical and hydraulic performance, and the cost of production as well, thereby making it a formidable analysis tool for design and revision. Moreover, the very important geotechnical property of clean soils and soils with fines (criteria for being well graded) can also be applied to gradation data, and values of shape factor can be obtained to establish it. For instance, well graded gravels exhibit shape factors between 0.37 and 1.29, while well-graded sands have shaped factors in the range of 0.37 to 1.00. These well-graded limits can be represented in the gradation chart. Moreover, RAMCODES also uses the gravel-to-sand ratio (G/S) as an alternative parameter to correlate gradations to UGM or HMA behaviour. G/S may be calculated from the shape factor and maximum diameter data using an analytical formula. Gravel-to-sand ratio is founded in another geotechniqueal principle at which curves for several G/S values can be represented in the gradation chart, as well (F. J. Sánchez-Leal, 2007).

However, these concepts of shape factor and Dmax were not considered in this study because they are limited by the continuous gradation. In addition, there is no clear correlation between these concepts and HMA performance in the study showed by (F. J. Sánchez-Leal, 2007).

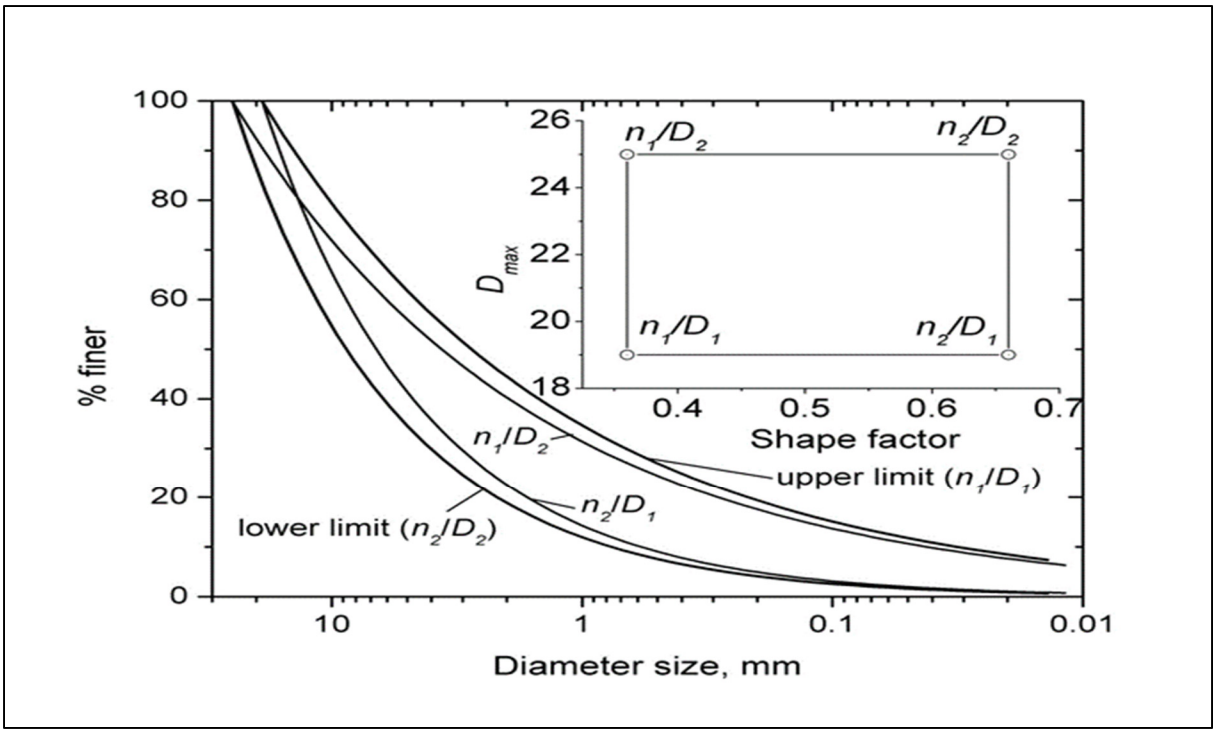


Figure 1. 2 Illustration of gradation-chart used in RAMCODES methodology taken from (F. J. Sánchez-Leal, 2007)

1.2.2 The Gravel-to-Sand ratio (G/S)

The gravel to the sand ratio (G/S) is a quantitative characteristic factor proposed by (F. J. Sánchez-Leal, 2007) in which gradation can be measured and referred to as a number. This gradation value depends on the relation between gravel proportion (particles larger than a sieve #4, 4.75 mm), and proportion (particles smaller than a sieve #4 but coarser than sieve #200, 0.075 mm). This factor attempts to basically simplify the traditional gradation nomenclature used in hot mix asphalt (HMA) in which the correlations among specifications and performance had remained unclear, thereby being difficult to understand why mixes grouped under the same specification presented different performance responses and mixtures classified as different exhibited similar performances.

For soil materials, designers might expect gravel material being stronger than sand material, and more permeable, at the same time. However, in HMA asphalt designers cannot say the same from, for instance, SP (Superpave mix) 9 mm to an SP 12.5 mm. It can be said that G/S

can allow the designer to anticipate the geomaterial behaviour readily from gradation data. This simplification produces robust correlations among performance like rutting, workability and permeability. The G/S factor surprisingly shows that permeability was independent of the type of asphalt binder under constant total voids and thickness layer. The study of (Sánchez-Leal 2007) based its finding on 11 highway projects through the United Stated (Cooley Jr, Brown, & Maghsoodloo, 2001) in which permeability test was performed using a variable-head device; measuring total void between 3 to 12 % for mixtures NMAS 9.5, 12.5, 19, and 25 mm. The G/S correlation results defined a critical value for permeability as the value of in-place total voids under which the permeability rate became maximal as can be seen in Figure 1.3.

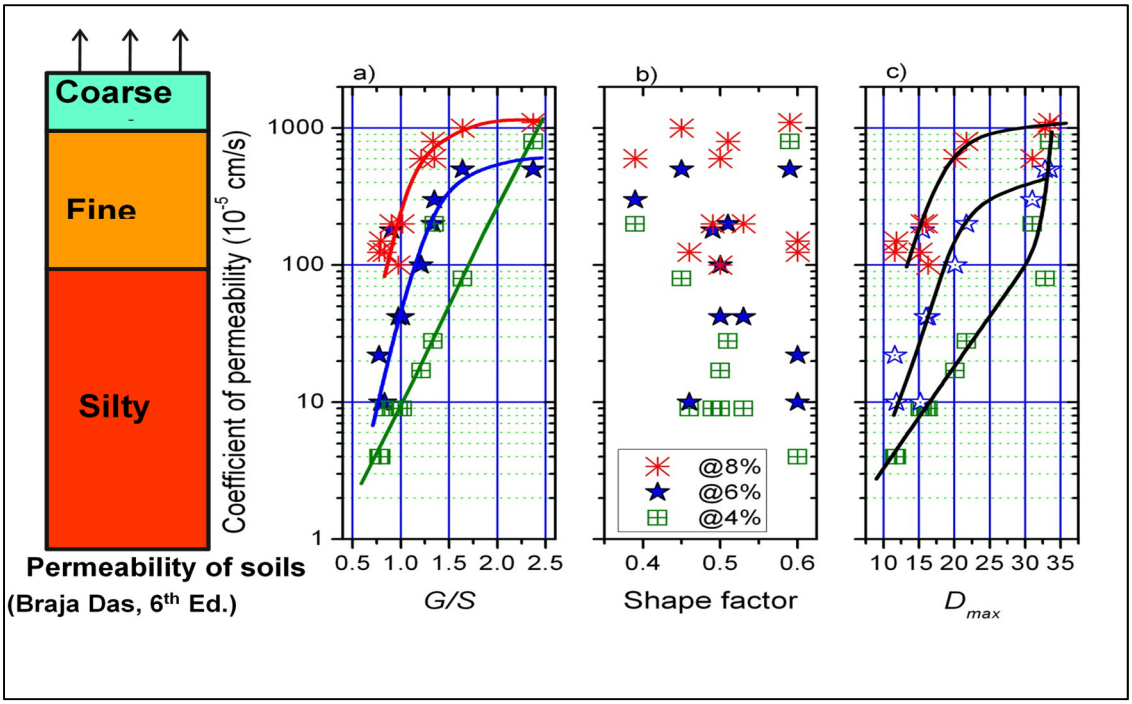


Figure 1. 3 Correlation between permeability with (G/S, Shape factor (n), Dmax and a material comparison permeability taken from Sánchez-Leal (2007)

Recently, some studies have adopted the G/S to correlate gradations with UGM properties. The way in which the study of (Xiao, Tutumluer, Qian, & Siekmeier, 2012) applied Ramcodes concepts developed for HMA to adequately evaluate UGM gradations, confirms the Ramcodes

postulated 1 (geomaterials). Moreover, the work of (Xiao, 2014) compared the G/S ratio with others like coarse aggregate ratio (CA), coarse portion of fine aggregate (FAc), and fine portion of fine aggregate (FAf) from Bailey's method to characterize aggregate blend.

Thus, the Bailey method's ratios were found statistically irrelevant, and the G/S ratio was selected as the main characteristic factor due to the strong correlations with the stability properties like resilient modulus and shear strength. This study was able to elucidate the characteristic of aggregate arrangement and porosity due to the strong effect of the G/S factor on the shear strength behaviour registered. The correlation between the G/S and performance displayed a curve in which values of G/S between 1.5 and 2 produced the highest values of peak deviator stress at failure. The Figure 1.4 illustrates the aggregate package arrangement among different states of G/S which small black dots represent the fines fraction, blue dots represent medium particles, and big grey dots represent the coarse.

Additionally, this study (Xiao 2014) elucidates the phenomenon of aggregate packing by using G/S to interpret the findings in which G/S equal 1.5 usually shows the best performance for UGM:

1) Largest G/S ratio (greater than 1.5) could reveal that there are no sand grains to occupy a portion of the voids between the coarse aggregate particles. Thus, shear or permanent deformation resistance is developed primarily by the interaction among gravel size particles (coarse) and could not be very stable depending on the grading of the gravel size particle distribution.
2) Optimum G/S (1.5) could mean that voids produced by gravel size particles are entirely occupied by the bulk volume of the sand grains, developing the condition of minimum porosity. Thus, producing more contact points among particles and the best performance regarding shear or permanent deformation resistance.
3) Small values of G/S (less than 1.5) could mean that more sand fractions exist until an optimal packing configuration is reached at the ideal state in which sand particles produce a disruption of coarse interaction, thereby reducing the performance of the aggregate mix.

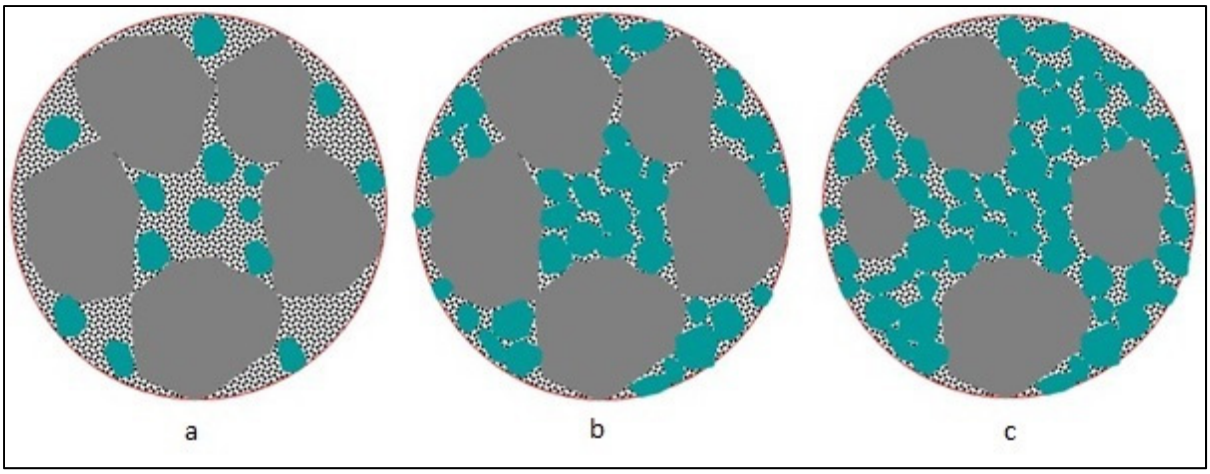


Figure 1. 4 The packing states of Gravel-Sand-Fines mixture with different Gravel/Sand: (a) Large G/S, (b) Optimum G/S, and (c) Small G/S taken from Xiao (2014, pp 160)

Furthermore, this understanding regarding aggregate packing applying the G/S concept was validated by the study of (Xiao & Tutumluer, 2017) in which using Discrete Element Model (DEM) to simulate the triaxial compression tests found that the highest numbers of contact points were achieved at G/S values near 1.5 specifically at 1.6 , thereby indicating a clear, well-connected contact force chain network compared with those with other G/S ratios. In addition, despite the angularity level of aggregate influence on a number of contacts and porosity, the values of G/S equal 1.6 constantly showed the more contact points, the small porosity.

To address the limitations of traditional gradation bands in order to classify a specific HMA, they are also based on: 1) the simplicity of the calculation implied, 2) the robust correlation found with the aggregate packing in which contact points were measured.

Thus, the gradation factor G/S was adopted as input values into the models developed in order to simulate the Rice test to obtain Gmm and the SGC test to obtain Gmm by focusing on the main objective to correlate gradations with volumetric parameters of the HMA.

1.2.3 The Characteristic Factor (Fp)

The *Fp* characteristic factor is a gradation operationalization approach initially proposed by (F. Sánchez-Leal, 2003; Sanchez-Leal Jucá, De Campos, & Marinho, 2002) to quantitatively

classify soil. This gradation characteristic factor is calculated as a linear product between finer-to-gravel ratio and liquid limit.

$$Fp = \frac{(1 + WL) * F}{(1 + G)} \quad (1.1)$$

Equation 1. 1 Taken from Sanchez-Leal Jucá et al., 2002

Where:

1) *WL* = Liquid limit.
2) *F* = Fine or percentage of passing 200.
3) G = coarse aggregate fraction of the aggregate retained on the 4.75mm (No 4) sieve. The larger the fine content, the greater the *Fp* factor. (e.g., *Fp* for *GM-GC* = 0.0-0.2).

Moreover, the study of (Sanchez-Leal Jucá et al., 2002) found for the soils material studied, robust correlation between: 1) *Fp* and Maximum Dry Density (MDD), and 2) *Fp* and Optimum Water content (OWC).

Furthermore, in an effort to control variations of HMA volumetric parameters in production plants, the work of (Sanchez Leal et al., 2002) suggested adapting the initial *Fp* calculation to HMA gradations. The initial *Fp* calculation quantifies a geomaterial based on gradations and specific surface of fines. However, the gradation of HMAs in general contains a low percentage of fines. Thus, the term $(1 + WL)$ was corrected by the value of 1.1.

$$Fp = 1.1 * \frac{F}{(1 + G)} \quad (1.2)$$

Equation 1. 2 Taken from (Sanchez Leal et al., 2002)

This quantification factor was used as input to develop the models under the hypothesis that G/S could not confer enough information to the DNNs model due to gradation discontinuities of some open mix HMAs in the dataset.

1.2.4 Gradation Index (GI)

Gradation index is an operationalization approach calculated as the ratio between the area of the retained area curve and the total area of an aggregate grading curve (A. Setiawan, L.

Suparma, & A. Mulyono, 2016; A. Setiawan, L. B. Suparma, & A. T. Mulyono, 2016; Setiawan, Suparma, & Mulyono, 2017). The equations 1.3 and 1.4 define the calculation; additionally, an illustration regarding retained area curve (the shaded area above the curve) is shown in Figure. 1.5.

$$GI = \frac{a}{A} * 100 \tag{1.3}$$

$$a = \sum_{i=0}^{n} \left(\frac{Sr_i + Sr_{i+1}}{2} \right) (T_i - T_{i+1}) \tag{1.4}$$

Equation 1. 3 and 1.4 taken from (Setiawan et al., 2017)

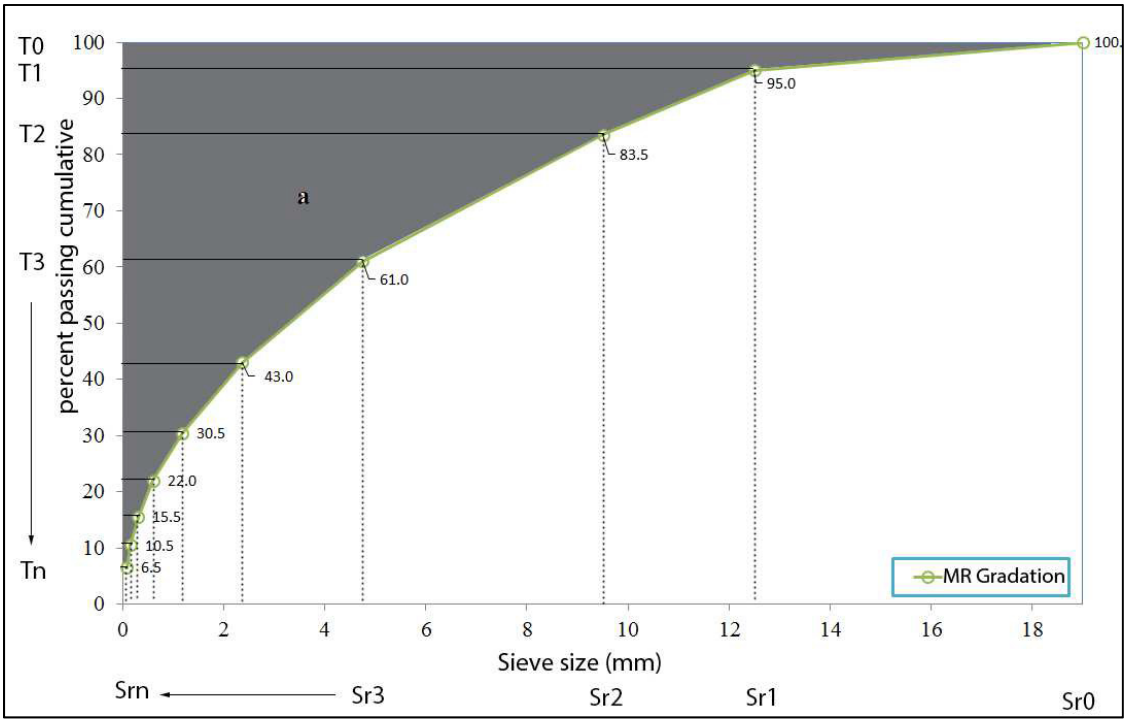


Figure 1. 5 Illustration of the retained area of gradation curve taken from (Setiawan et al., 2017)

Gradation index (GI) concept was adopted in this study to quantify the aggregate gradation because GI considers all the sieve size in contrast to the Fp, G/S which only taking into account specific sieve size.

1.2.5 Total Granulometry (TG)

The TG is a gradation quantification factor which is obtained by the sum of sieve passing from 20 mm to 80 microns (Quebec, 2019) . This gradation operationalization is used in the province of Quebec in Canada, as a simple way to control the quality of HMA in asphalt plants. This gradation factor allows the quality control inspectors having an idea if the gradation is coarser or finer.

$$\mathrm{TG} = \sum \mathrm{passignform\ sieve\ 20mm\ to\ 80\mu m} \tag{1.5}$$

Equation 1. 4 taken from (Quebec, 2019)

1.2.6 Discussion Gradation Factors

The variation in the traditional specification band does not allow researchers quantifying the aggregate gradation, thus making difficult understanding changes in the behaviour of the aggregate gradation (Setiawan et al., 2017). For this reason, the operationalization as a number in a continuous scale of aggregate gradations should be applied, thereby achieving correlations between response parameters with gradations (F. J. Sánchez-Leal, 2007).

Several quantification factors have been found in the literature to correlate aggregate gradations with HMA performance (workability, rutting, Marshall's stability, permeability, tensile strength). Moreover, researchers have been proposed correlation between gradation factors and HMA volumetric properties; for instance, the study of (Shen & Yu, 2011) has proposed a correlation between aggregate structures and volumetric properties to predict VMA.

However, not only any direct correlations between gradations and the HMA volumetric properties (Gmm and Gmb) have been found in the literature to optimize the HMA design, but also researchers still must find a way to correlate HMA volumetric properties with these four gradation factors (G/S, Fp, GI, TG). This study explores the use of these factors to evaluate the

relative importance of these factors on the HMA volumetric phenomena of the Gmm and the Gmb.

1.3 Hot Mix Asphalt (HMA) Design

The term hot mix design is referred to the procedures necessaries to obtain a stable aggregate blend bounded by asphalt in which the main performances like durability and security are achieved. Several types of HMA with different approaches and limitations can be played to try to design a bound material, thereby achieving these two-principal performances. These designs could be classified by volumetric and performance results.

1.3.1 Volumetric HMA Design

This HMA deign is an approach in which a state of voids is reached by compacting specimens under the impact like Marshall's method or kneading action like Hveem or Superpave method. This state of voids has been empirically adopted by consensus in the field of HMA due to performance like rutting, stiffness, or fatigue have been indirectly correlated with the HMA internal voids. Some studies have found limitations by using this kind of approach, thus evidencing that the major benefit is the simplicity of running tests like 1) a compaction test and 2) a suction test to find a hypothetical maximum compaction in which no voids can be present, thus considering only the mass and volume of materials (aggregates and asphalt binder) to obtain the volumetric properties Gmb and Gmm and establishing the state of voids. This simplicity enables the quality control in asphalt plant despite limitations regarding to performance correlations.

1.3.2 Balanced Mix Design (BMD)

The Balanced Mix Design (BMD) is a type of asphalt mix design in which basically the state of voids must be validated by running performance tests on properly conditioned specimens or directly designing HMA under performance approach tests by which mix aging, traffic,

climate, and location within the pavement structure are considered. (West, Rodezno, Leiva, & Yin, 2018)

This design concept typically includes rutting test and a cracking test, thereby evaluating the mixture resistance to these two forms of distress (Yin, Taylor, Tran, & Director, 2020). Moreover, this design balance concept contains three approaches as can be seen in Figure 1.6 . The first, the Volumetric Design with performance verification; the second, Performance-Modified volumetric design; and the third, Performance Design.

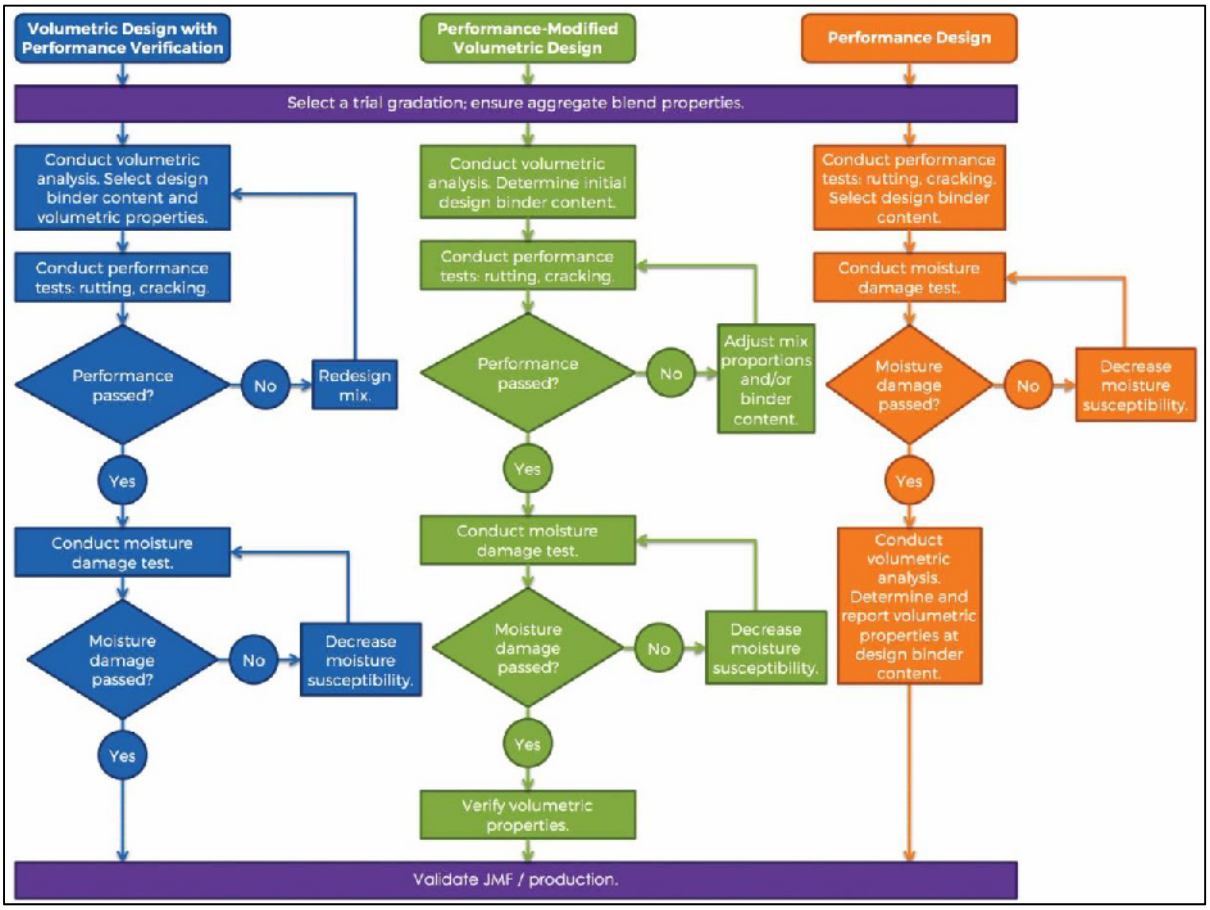


Figure 1. 6 Diagram of the three BMD approaches taken from (Yin et al., 2020)

1.3.3 Discussion HMA Design

Despite the advantage of volumetric design in the quality control of HMA in asphalt plant, several researchers have shown that different compaction techniques reflect in different states of voids. These differences regarding the states of voids when compaction technique is changed, make it difficult to understand the aggregate packing phenomenon, thus the correlation between performance and the HMA volumetric pro perties becoming unclear. For example, the work of Mollenhauer and Wistuba (2013) compared laboratory procedures (impact, gyratory, compactors with pneumatic tires or with smooth steel roller) for compacting HMA, based on the triaxial cyclic compression test (TCCT) under the European specification EN 12697-25. This study showed that different compaction procedures reacted in a different way during TCCT (three-dimensional (3D) stress).

The work elucidates that the external diameter of gyratory and impact briquettes could contain more voids attributable to the mould contrasted with the cores cut from compacted slabs or the field. Thereby, the axial strain rate measured for gyratory and impact compaction was relatively low in which strain rates were approximately 0.05 μm/m per load cycle marginally affected by the void content of the specimens. The work explains that probably the mould avoids horizontal movements of aggregate particles when vertical compaction force is applied through impact and gyratory compaction, thus packing the aggregates only vertically and forming a skeleton weakness of spreading horizontal loads.

In fact, the study exposes that the confining stress enables to exclude radial deformation or to produce negative radial strains, thus reducing the diameter of the specimen. However, the specimens compacted by laboratory roller procedures showed a significant influence of the void content on the radial strain rate, in the same way as for the axial strain rate.

In summary, this research concludes that:

1. laboratory roller-compacted specimens are more influenced by the content of voids regarding the 3D deformation behaviour than specimens compacted by impact or gyrations which showed having lower susceptibility to void content variation, thereby resulting in higher values of the stiffness compared with roller-compacted mix.

2. specimens compacted by laboratory roller correspond the better with the 3D deformation behaviour of field-compacted asphalt mix than specimens compacted by impact or gyratory; despite, the moulds of laboratory roller-compacted specimens constantly induce higher resistance to deformation and higher stiffness compared with field-compacted asphalt mix.

This unclear correlation is the main reason by which researchers have adopted the modern approach of Balanced Mix Design (BMD) in which complementary to the Superpave mix design, the BMD involves performance properties validation of the final mixture rather than specifying volumetric properties and formula components, thereby allowing engineers apply a free design in which the use of innovative materials and technologies for asphalt mix could be accepted by agencies (Yin et al., 2020). However, the study of (Yin et al., 2020) found important differences in the performance test results of the BMD mixes between mix design and production. Thus, a procedure for QC and acceptance testing of BMD is proposed and discussed. This additional procedure contrasts with the simplicity of volumetric design (Superpave) regarding the quality control in asphalt plant despite the weaknesses in correlating performance with volumetric properties.

1.4 Artificial intelligent and Machine Learning Algorithm Concepts

1.4.1 Artificial Intelligent (AI)

Humans have been trying to optimize activities to facilitate the life shortly after they appear in the world. This natural act has been achieved due to the humans have an important and specific quality different to other creatures in this world. This quality is the Intelligence which permits them (humans) to identify, understand, speculate, expect, and control things that can be more complex than themselves (Russell, 2010). Thus, devices to help humans to perform tasks can usually be seen around when cars, bikes, airplanes help to transport individuals or calculators and computers allows engineers produce better designs. However, these common devices were no conceive by their creators with the capacity to think (having intelligence).

In facts, the extent to which researchers have formalized the language of logic reasoning began in 1847 with George Boole, almost a hundred years after Alan M. Turing wrote about Turing-machine opening the door of artificial intelligence (AI) (Buchanan, 2005). In effect, having intelligent devices or machines were deep in the imagination of the people until the advance computer appears in the 1940s when Alan Turing's laboratory in Manchester, the Moore School at Penn, Howard Aiken's laboratory at Harvard, the IBM and Bell Laboratories demonstrated the computer calculating power (Buchanan, 2005).

Thereby following these steps in 1943, Warren McCulloch and Walter Pitts developed the artificial neurons model which simulating connection information among neurons of the human brain. This contribution of McCulloch and Pitss helped to develop in 1944, the theory of decision by J. Neumann and O. Morgenstern; only a year after in 1949 Donald Hebb showed the possibility of learning when changes rules of the connections of the artificial neurons were made. Indeed, it was in 1950 when Alan Turing proposed concepts regarding the possibility of a machine could think by programming an electronic computer to perform intelligently (Buchanan, 2005).

Finally, in 1951 Marvin Minsky and Dean Edmonds created the first neural computer (Benko & Lányi, 2009). These concepts suggested by Alan Turing were explained under the Turing test approach, also known as imitation game, in which instead of elucidating whether machines can think or not, it was more important to know whether machines could pass a behavioural intelligence test.

This Turing test is based on evaluating whether a communication (via online typed messages) could be achieved with a program software for five minutes. The interrogator (human) must guess if the communication was with a program or a human. The model program succeeds if the interrogator is confused 30% of the time (Russell 2010). It seems that the results of this test could depend on human skills to evaluate the answers.

However, even if programs need to deal with a sophisticated judge, nowadays the world lives in the age of big data in which computer has the capacity to collect vast quantities of information impossible for a human to manage (Anyoha, 2017). Thus, it is common now in this age, being confused with the modern customer services, referred to intelligent virtual assistant, provide by tech companies like Apple (Siri), Microsoft (Cortana), Amazon (Alexa),

and Google-assistant in which it is not clear for many persons whether the communication is with a computer or not.

Furthermore, the fast development of computers has allowed researchers to apply the Artificial Intelligent (AI) approach to rapidly solve problems that are rationally complicated for humans; nevertheless, relatively simple for computers when problems can be described by a list of formal, mathematical rules (Goodfellow, Bengio, & Courville, 2016).

For this reason, some researchers merely define the AI as the model selected to solve a problem (Heaton, 2013). Nonetheless, this advantage over humans depends on the full description of the information by a concise list of formal rules that can be provided by the programmer in advance. For computers is difficult to obtain knowledge from informal data as the humans intuitively do by performing tasks like merely identifying a specific food by taste-smell connection.

To address this limitation, researchers developed an AI system with the capability to obtain their own knowledge in which the raw data collection allows the system identifying patterns. Thus, this AI system capability is referred to Machine Learning (ML) which enables computers to deal with problems regarding knowledge that appear subjective of the real world to draw valid conclusions (Goodfellow et al., 2016).

The technological development in the field of computer science has grown over time as shown in Figure 1.7. For more details on the AI definition see (Goodfellow et al., 2016; Russell, 2010)

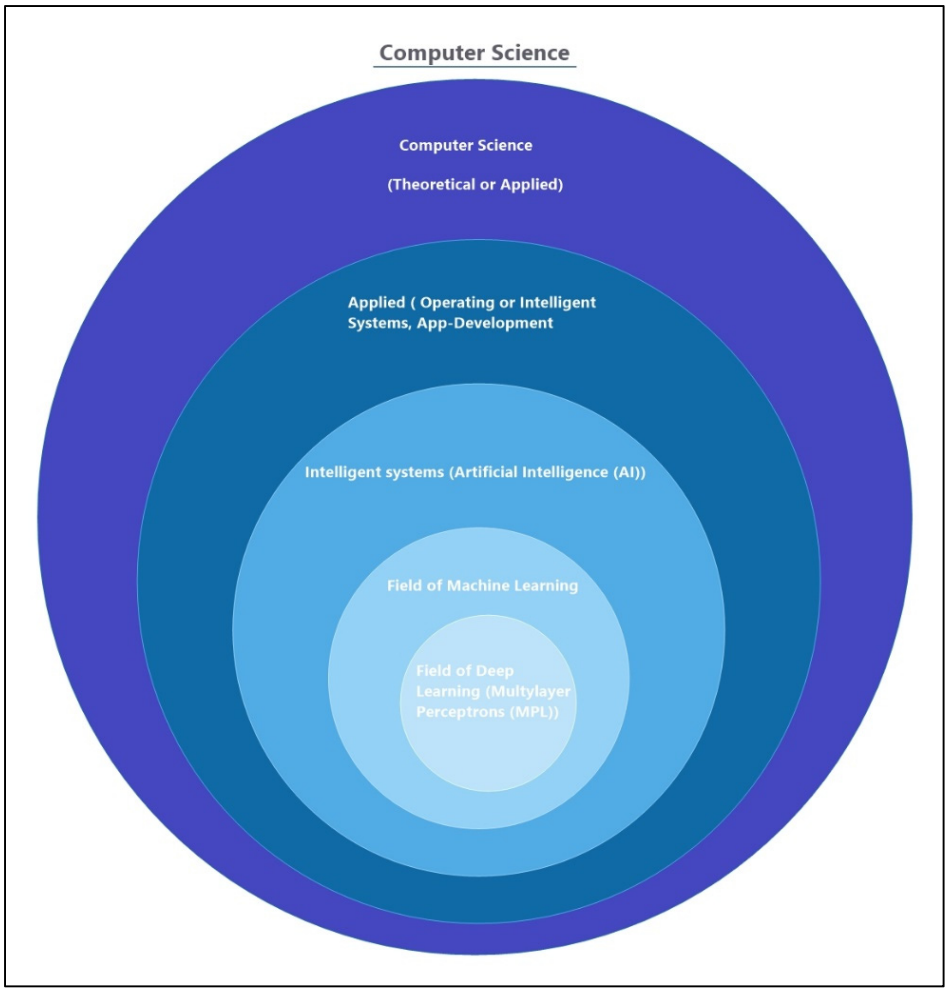


Figure 1. 7 A rough illustration of the computer science development

1.4.2 Machine Learning

Machine Learning (ML) is the concept used which an AI system allows the model adjusting chores, learning, and increasing experience from the data without any human intervention. Thus, a computer program is developed which enables to access data and learn by himself (Goodfellow et al., 2016). Then, ML could be defined as the way by which algorithms and techniques are used by computers, thereby producing by themselves, decisions, and predictions from an accessible data.

However, Goodfellow et al. (2016) explain that this AI computer program depends on the data representation to adequately perform, thereby dealing with this dependence is a common fact

in the field of computer science. This data representation challenge can be solved by quantifying (operationalization) the informal information by which abstract concepts are converted into measurable observations, thereby collecting systematically data that are not directly observed could be achieved by designing the right set of features to extract information. This set of features as can be seen in figure 1.8 can provide enough information to the machine-learning algorithm for a task.

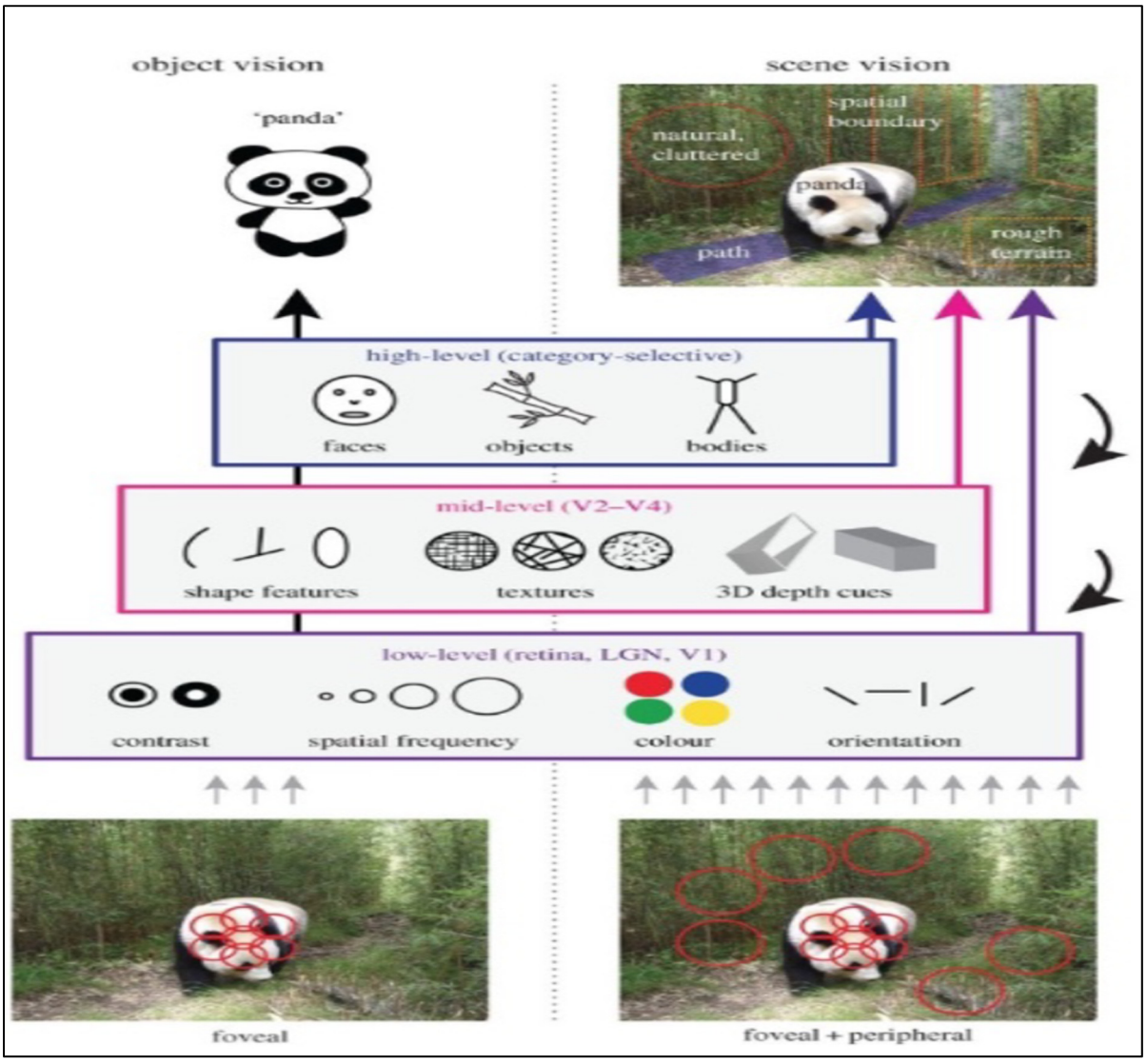


Figure 1. 8 Low-Mid-High Level Features. This sketch shows how the different class of features can be used to address different problems in which the main concept provided by an image is maintained by the contribution of each stage of the representation. Taken from (Groen, Silson, & Baker, 2017)

For example, a possible feature for a simple drink, fruit or food identification (natural straightforward for humans) could be an estimation of the pH (acidity), thus allowing the algorithm having a good idea whether the drink, fruit or food is sour or sweet.
Nevertheless, the problem remains as to how represent data if each piece of data observed is affected by other factors of variation (hidden variables) which is the case in the real world of artificial intelligence applications. For instance, the pH of two different fruits like tangerine and grapefruit could be similar if they are compared at different age of maturity because the pH or acid levels of the fruit decrease as the fruit ages. In addition, extracting abstract features from raw data is complicated and only possible under practically human participation by understanding of the data (experience).
Recently researchers have made significant progress in overcoming the difficulty regarding representation data, by developing the deep learning (DL) concept in which an artificial neural network (ANN) referred to multilayer perceptron (MLP) or deep neural network (DNN), has complex connections between layers and more neurons (Abiodun et al., 2018), thus allowing computers developing complex models using simpler models by which the data is directly observed and the representation data of the complex model is obtained in terms of simpler representations (Goodfellow et al., 2016).

1.4.2.1 Machine Learning Techniques (MLTs)

Machine learning techniques (MLTs) are classified into three categories: 1) supervised, 2) unsupervised and 3) Reinforcement learning. A sketch of the most commonly used algorithms is shown in Figure 1.9. It should be noted that the Artificial Neural Network is bold (blue) to indicate that this approach is a supervised learning algorithm which was adopted as a method to developing prediction models in this study. For more details on the ML definition and techniques see (Goodfellow et al., 2016; Haykin, 2009; Russell, 2010).

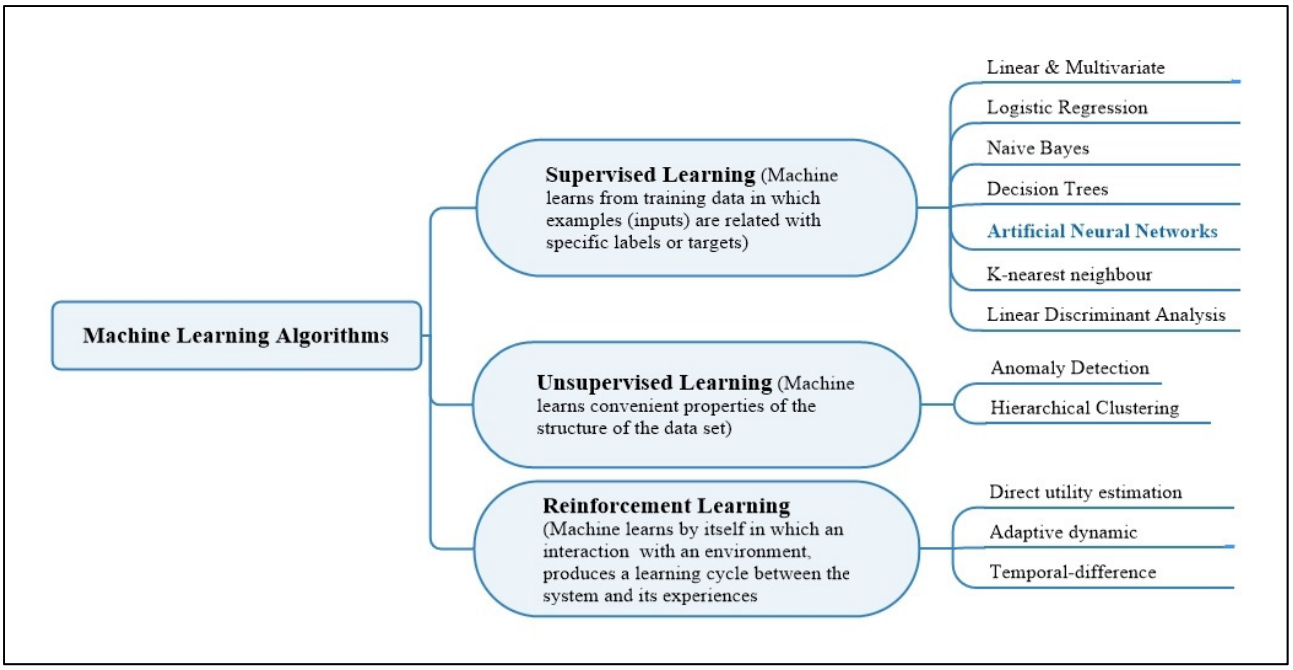


Figure 1. 9 Summary illustration of the types of machine learning algorithms

The work of (Liu, Zhao, Ju, & Shi, 2017) has summarized in three essentials steps how researchers should apply MLTs in materials science. Figure 1.10 shows in general these three steps as following:

1) Sample construction: This is an important step because in materials science, simulations and experimental measurements provide the raw data (observed). This raw data usually has some problems regarding noisy, inconsistent, and incomplete values due to human error, thereby cleaning data procedures are strongly recommended to be performed when constructing a sample from the original data.
2) Model building: Complex links commonly exist between the conditional factors and the target attributes in material science research in which traditional methods have difficulty in handling them. MLTs allow researchers using examples of a target function to discovery the coefficients, thus developing a certain mapping function in which the target function is properly approximated. For this reason, it is crucial exploring and using a particular set of non-linear or linear functions to clearly observe a relationship between input data and output data.

3) Model evaluation: The performance of a model is evaluated by the extent to which the model can achieve a good performance on unseen data. This evaluation is usually made by examining the generalization errors of models by means of calculation-based tests and use the results to select the best one. Thus, testing data are required to assess the capabilities of models on an unknown dataset. Subsequently, the error produced by this test could be taken as an approximation of the generalization error.

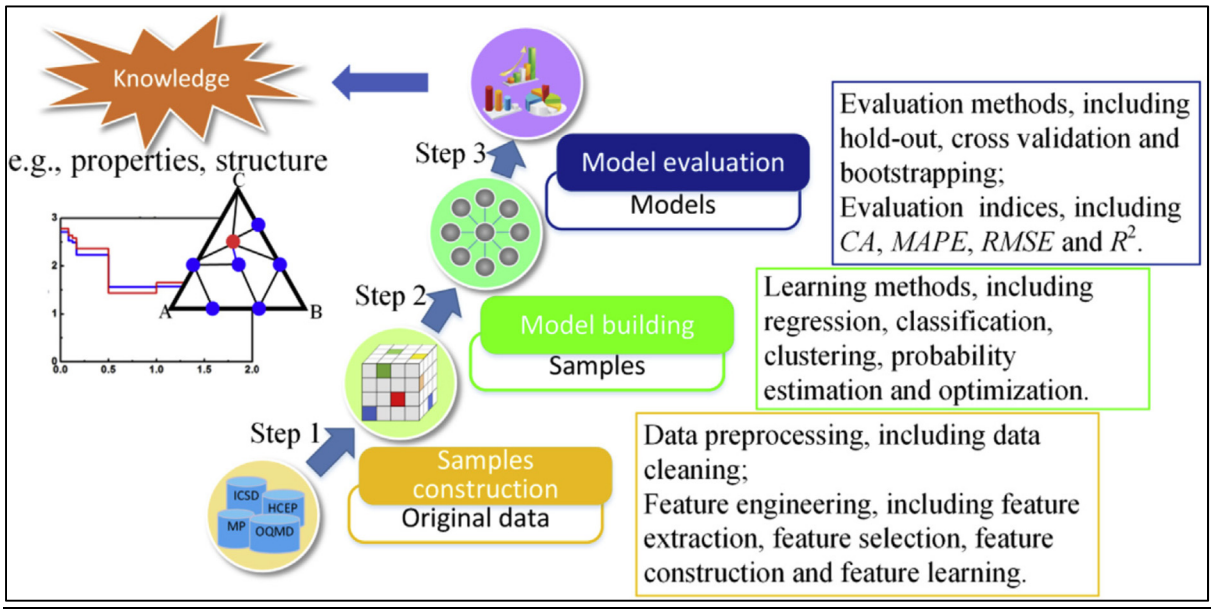


Figure 1. 10 The general process of machine learning in materials science. Taken from Liu, Zhao et al. (2017)

1.4.2.2 Supervised Learning Algorithms

Supervised learning is the term used to define an algorithm that enables models to acquire experience by labelling the input features with the associated real value or target (outputs) of the dataset used for training. This MLT calculates the differences between the model outputs and the consequent ground truth values to build the objective function, thereby training process of the model is applied with the intention of minimizing the constructed objective function (Huyan, 2019). An example of supervised learning is when a person tries to learn how to make a cake under the supervision of someone (mother or grandmother) whose can correct each step

of the cooking process. In effect, the person can verify if each step has been performed as the same way as the inspector did and if the result (cake) looks like the cake made by the inspector, thereby adjusting the temperature of the oven (weights) using the same mixing technique (neural architecture) and same ingredients (inputs) to obtain similar results.

1.4.2.3 Unsupervised Learning Algorithms

Unsupervised learning is a machine learning method referred to the technique that classifies the datasets avoiding labelling the datasets before the training process. This MLT is based on merely as a set of input signals for which an abundant supply is available, thus learning the intrinsic knowledge only based on input instances could be achieved by the algorithm even though no explicit feedback is supplied (Haykin, 2009; Huyan, 2019; Russell, 2010) . The unsupervised learning methods create clusters, in the dataset containing many features, in an area, thereby identifying similarities among the objects for classifying unknown patterns (Abiodun, Jantan et al. 2018).

For instance, a baby used to watch African animals' pictures with his father (lions, elephants, giraffes, etc). The baby has no idea that in the real world there are many lions, elephants, and giraffes. After several months, all family went to the zoo, the baby has not seen any of the animals which live in the zoo earlier. However, the baby is able to recognize them, as a lion, as an elephant, and as a giraffe through a set of features such as two ears, four legs, a tail, fur, whiskers, trunk, etc; even whether the lion has a different colour (white) or the elephant is a cub.

1.4.2.4 Reinforcement Learning

Reinforcement learning is the learning system which executes a series of actions and it learns from the response (rewards or punishments) of the environment (Haykin, 2009; Russell, 2010). For instance, a person that is learning how to make a cake by himself without the video reference from YouTube, neither a recipe with instructions procedures nor any helper adviser verifying procedures. Then, he will start by turning on the oven, mixing eggs and butter in a

bowl, then adding sugar and flour. He continually practises cooking cakes steps in a large open field and gradually master the cooking technique skillfully by making sense of the data by evaluating whether the cake is well cooked or overcooked or it needs to adjust cooking time and temperature without a guide. Thus, the person will be able to learn which ingredient has more influence on the cake and how to optimize the technique to obtain a good cake, whether the cake is easily sold (reward) or the cake is not sold (punishment), by losing money and time effort.

1.4.3 Artificial Neural Network (ANN)

The ANN is an approach of machine learning (ML) motivated by the basic findings of neuroscience under which the mental activity of the biological nervous system function is mainly an electrochemical activity of the brain cells referred to neurons (Russell, 2010). The ANN is composed of nodes (neurons or units). These nodes receive information from inputs (features) presented at the visible layer, which contain the variables observed. These inputs are multiplied by numeric weights. The weights determine the strength of the connection in which the more relevant the input, the larger the weight. Then, an activation function is triggered by processing the inputs, thus producing an output that will be the input of other neurons into the network (Al-Mosawe, 2016). In summary, the nodes receive input from previous nodes and deliver output to subsequent nodes as can be seen in the Figures 1.11 and 1.12.

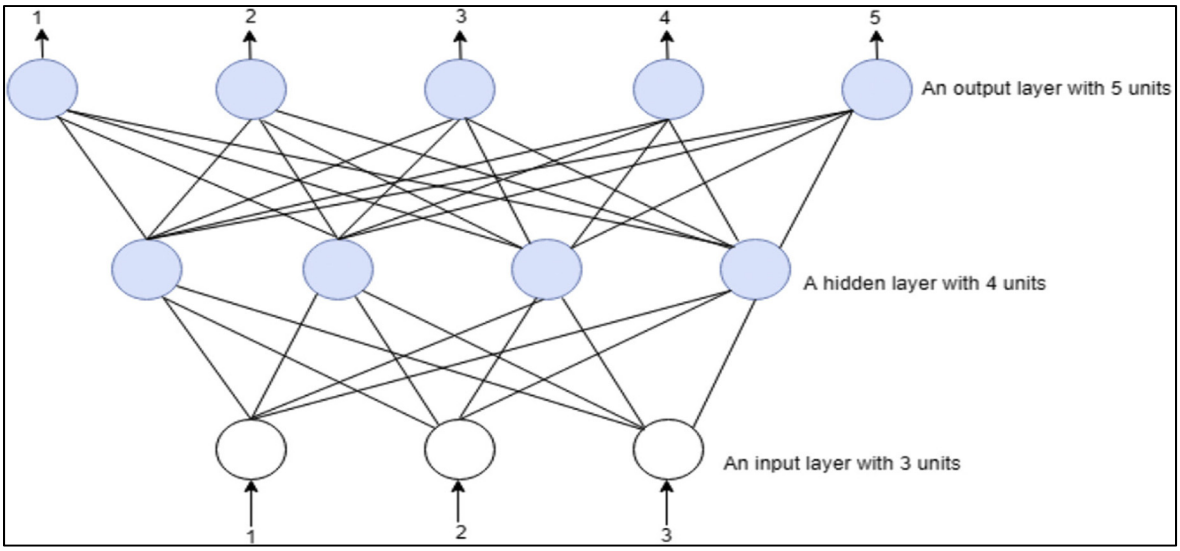


Figure 1. 11 Two-layered feed-forward neural network taken from (Abiodun et al., 2018)

This process in referred to feed forward propagation (Abiodun et al., 2018) because the neural connection is in one direction. The feed-forward networks are structured around a chain of functions (layers) by which the first layer of the network contains the first activation functions into the nodes, the second set of functions is referred to the second layer, thus giving the depth of the model: the extent of the chain of the longest path from input to output. Finally, the last layer is referred to the output layer. In addition, because the required output is not achieved during the training process after every layer, these layers between the inputs and the last layer (output layer) are referred to hidden layers as can be seen in the Figure 1.11 (Goodfellow et al., 2016).

ANN architectures applied to solve problems like regression, and function approximations are usually feed-forward multi-layer perception (MLP) in which the training process is based on adjusting the weights and biases of the connections to minimize the error between the network outputs and the real values. For this reason, the Mean Squared Error (MSE), Mean Absolut Error, and Mean Root Squared Error (RMSE) are often used as a performance indicator for the network, and the back-propagation gradient descent method is adopted for training the ANNs, thus adjusting simultaneously the weights and biases (Huyan, 2019). Additionally, the term DNN is used to refer to as an ANN when it has more than three layers or complex multilayers as the MLP. An explanation regarding metric error used in this study is shown in chapter 2.

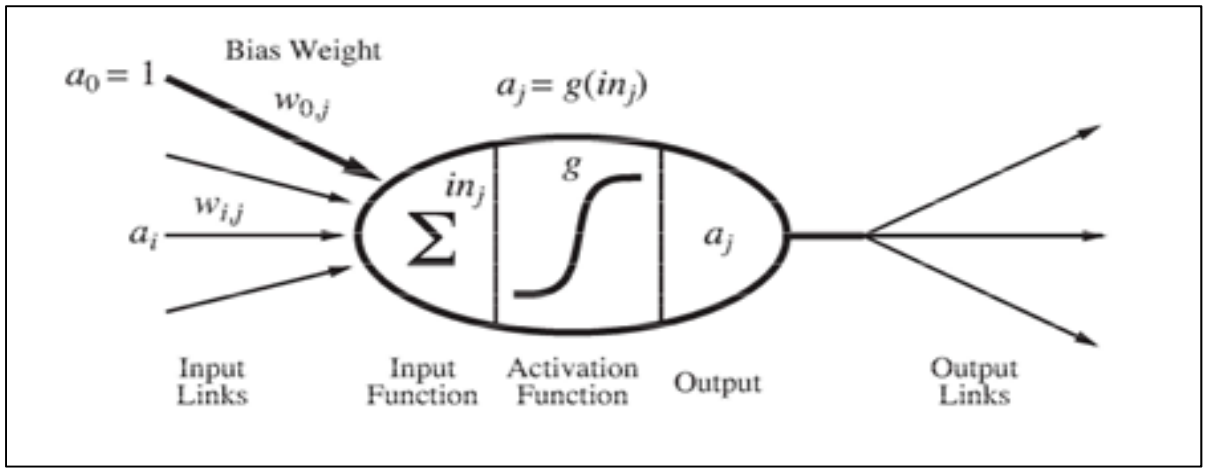


Figure 1. 12 Mathematical model for a neuron. The unit output activation (ai) equals to $(g(\sum_{i=0}^{n} wij * ai)$, where ai is the output activation of unit i and wij is the weight on the link from unit i to this unit. Taken from Russell (2010, page 728)

The accuracy of the ANN depends on the appropriate architecture and hyperparameter selection (number of layers, nodes, learning rate, dropouts, batches, etc.) with appropriate activation functions and model training implementations (Huyan 2019). However, for ML algorithms the challenge is to develop a model which performs adequately on new data (unseen inputs), instead of just those on which the model was trained. This capacity to adequately perform on unknown data inputs is referred to as the generalization (Goodfellow et al., 2016). In fact, selecting an extensive neural network architecture could develop the model which contains too many parameters, thus being able to memorize all data examples, and increasing its capacity that could not effectively generalize to data samples that have not been seen before during the training (Russell, 2010).

In contrast, models with low capacity (small architecture) may have difficulties to fit the training set (Goodfellow et al., 2016). In effect, the study of Goodfellow et al., (2016) explains that evaluating how a ML algorithm performs, is based on the ability to produce a small error in the training and reducing the gap between training and test error. In effect, this performance is evaluated, by applying the concepts of underfitting and overfitting. Defining Underfitting when the model struggle to achieve a sufficiently low error value on the training set and

overfitting when the gap between the test error and training error is too big, as shown in Figure 1.13 where the training error (blue curve) and generalization error (green curve) have different behaviour in which increasing the capacity indicating that:

1) the training error curve decreases. However, the gap between training and generalization error increases.
2) Both error curves (training and generalization) are high at the beginning, but both decrease just until a point (Optimal Capacity) when generalization error curve increase, thus dividing the chart in two zones (Underfitting at the left of the graph and overfitting at the right of the graph).

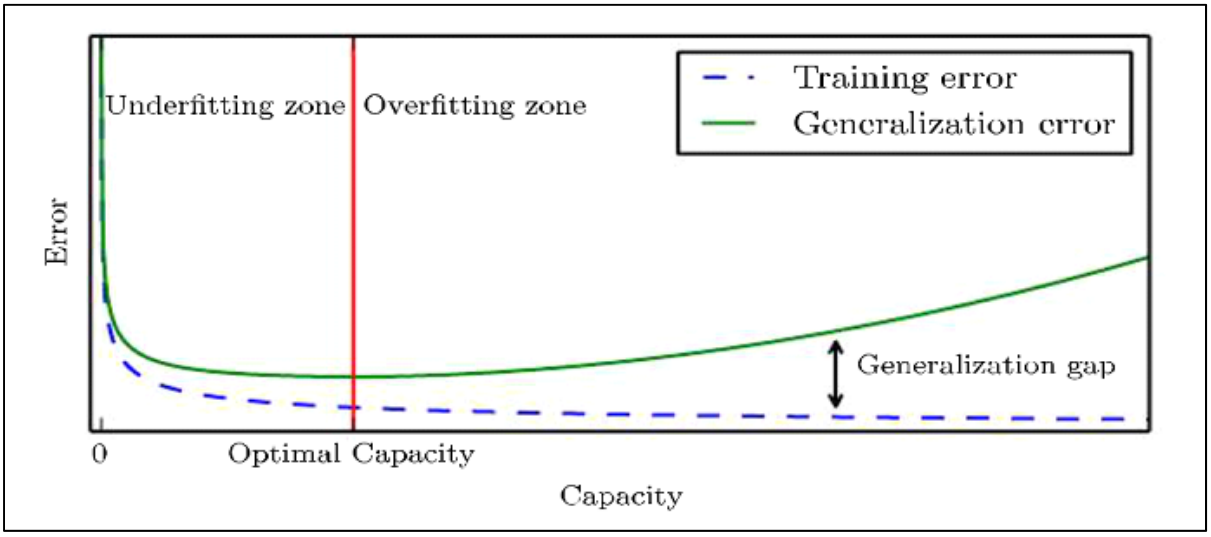


Figure 1. 13 Typical relationship between capacity and error. Taken from Goodfellow et al. (2016, page 113).

For the purposes of this work, the term of Deep Neural Network is an artificial neural network (ANN) with more than one hidden layer in which complex connections are developed between layers and neurons. For more details on ANN definition see (Goodfellow et al., 2016; Haykin, 2009; Russell, 2010).

1.4.4 Cross Validation

The cross-validation is a procedure which consists mainly of randomly selected different subsets by training and testing computation repeatedly. The Figure 1.14 shows that the partition of the original dataset is formed by splitting it into k non-overlapping subsets. Thus, the estimation of the test error can be obtained based on the average test error across k trials (Goodfellow et al., 2016) .

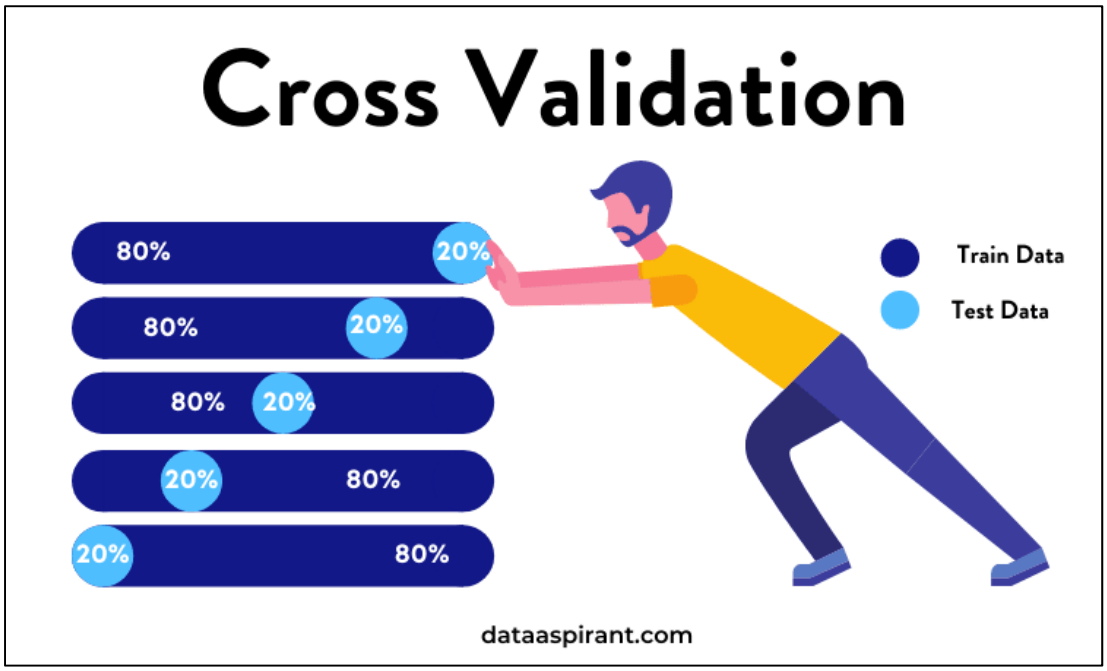


Figure 1. 14 Illustration of a cross validation taken from (Arif, 2020)

This cross-validation concept is applied when a small test set produces statistical ambiguity in the average test error estimation, thereby remaining unclear if one algorithm generalizes better than another for the same task. However, this limitation is not a serious issue if the dataset contains hundreds of thousands of examples (pairs) (Goodfellow et al., 2016).

1.5 Machine Learning Approach in Hot Mix Asphalts

Several researchers have been applied MLTs to develop models that can adequately predict HMA performance values like fatigue, permanent deformations, stiffness. For instance, the study of (Feipeng Xiao, 2009) elucidated that predicting the fatigue life was more efficient applying the ANN techniques in modified mixture than traditional regression-based prediction models. This work applied ANN concepts to predict the stiffness behaviour of rubberized asphalt concrete mixtures with reclaimed asphalt pavement (RAP). The ANN structure was based on 5 inputs variables such as 1) the viscosity, 2) creep stiffness values of the asphalt binders, 3) optimum binder content of the mixture, 4) indirect tensile strength (ITS), and resilient modulus (MR) and one output variable the stiffness (Sm) by which 84-stiffness examples datasets for the specimens were selected. The models were trained at 5 °C and 20 °C of temperature for ambient and cryogenic rubberized mixtures, thus developing an overall of four (4) ANN models.

The data (84 example) was split in 63 examples for the training data, and the other 21 examples were used as the testing data set. However, the ANN architecture of the models is uncertain about the number of nodes per hidden layer and the number of hidden layers into the network. These models apparently showed robust correlation between inputs and the output (Sm) represented by a testing R^2 of 0.57 at 5 °C and 0.71 at 20 °C (ambient rubber); for cryogenic rubberized mixtures, the correlations were higher with R^2 of 0.89 and 0.88 at 5 °C and 20 °C, respectively. Nevertheless, any error (training and testing) curves were displayed to assess whether the models could be overfitted or not. Moreover, despite the study established a validation of ANN models with examples from other projects, it remains unclear if the sensitive analysis was made on an overfitted model because the RMSE increased in 265 % from 1.21 to 3.25 at 5 °C temperature in the validation test.

Some studies have also modelled performance like the Marshall Stability (MS) by using ANN concepts. For example (Ozgan, 2011) developed a simple ANN architecture which contains five (5) input variables (volume of voids, unites volume weight for saturated surface and for dry air, temperature, and exposure time); one (1) hidden layer; and one (1) output variable the MS. This study collected 60 samples which were divided in 52 samples (80% of all samples)

for training, and the residual data (13 samples 20% of all the experimental results set), were selected for the test. Apparently, the study shows an adequate performance of the model developed with R^2 of 0.933 for testing. However, neither error curves (training and testing) were exposed, nor validation procedures with unseen data from other projects were made. This validation step seems to be basic according to the literature of MLTs to elucidate whether the phenomenon was properly generalized, or the model was overfitted.

Furthermore, researchers recently have made significant progress in overcoming the difficulty of traditional regression using ANN when many variables must be considered to accurately assess the phenomenon. For example (Al-Mosawe & Thom, 2016) developed new volumetric ratios among the coarse particles and fine fraction avoiding Bailey's weakness, thus establishing robust correlations between permanent deformation (rutting) and aggregate packing volumetric parameters. However, despite the models developed showed strong correlations among input variables and the output (rutting), the issue of generalization apparently remains because the study has certain limitations by displaying error (training and testing) curves and any validation was made with examples unseen by the models. Thus, the risk of conducting a sensitivity analysis in models overfitted increase.

Moreover, other researchers have been assessing the HMA volumetric approach using ANN by modelling air void content in aggregate mixture (VMA). For instance, the study of (Zavrtanik, Prosen, Tušar, & Turk, 2016) showed that the asphalt binder content is an influential factor which improves the model's performance. The study developed models for 7 types of HMA independently by using a dataset for each asphalt mix. These datasets contain between 176 to 381 pairs; additionally, other models were developed using a dataset built with the 7 datasets to achieve 1634 pairs. The results showed that linear models had better performance than ANN models when specific asphalt mixture was evaluated. However, ANN detected hidden relationships between data when dataset of all asphalt mixtures together was used, thereby being more effective than the linear model. These changes in the ANNs performance could be attributed to the quantity of data. Furthermore, despite datasets could be diversely represented, the problem remains as how to evaluate the ANN performance if neither learning curves were displayed, nor validation process using unseen data by the model was made.

Furthermore, the work of (Gong, Sun, Hu, Polaczyk, & Huang, 2019) applied the MLT referred to random forests to explore the impact of asphalt mixtures properties on pavement performance. This work used data from the Federal Highway Administration (FHWA) specifically from the Long-Term Pavement Performance (LTPP). The study used a total of 404 observations for cracking, 460 observations for rutting, and 443 observations for the roughness (IRI) which shows a sensitive analysis through the relative importance of the mixture properties input variables to HMA performance in the field (Alligator cracking (AC), Wheelpath longitudinal cracking (LC), Transverse cracking (TC), Rutting, Roughness (IRI). This relative importance exploration revealed that AC is very sensitive to gradation mixture, being the aggregate volumetric properties as Gsb one the most important inputs to AC, TC, and rutting phenomena. However, the results show that binder content had less impact or relative importance to the same phenomena.

The work of (Ceylan, Kim, & Gopalakrishnan, 2007) developed models for predicting HMA dynamic modulus using the Artificial Neural Network (ANN) methodology. These models were trained and tested using the same database of 7,400 records used in the development of revised |E*| Model of (Bari and Witczak, 2006). In addition, the ANN architecture used the same eight input variables of the 1999 and 2006 version |E*| of Witczak predictive equations. Thereby developing two models: 1) the ANN 1999 and 2) ANN 2006) with the |E*| as output variable in both the ANN models. The dataset of 7,400 records was split into two subsets, 6,900 data points for training and 500 data points for the testing. The work displays learning curves based on MSE progress for several network architectures with two hidden layers, thereby evaluating the performance generalization of each model. However, despite the study seems to avoid overfitting problems assessing the learning curves, it remains unclear if the dataset is focused on only one section of this spectrum where models tend to be responsive only on a small part of the spectrum because the validation process with unseen observations by models was not made in this study.

1.5.1 ANN to predict HMA volumetric parameters

Some researchers have seen the importance of optimizing the traditional volumetric mix design applying ANN. For instance, the study of (Ozturk & Kutay, 2014) made an important contribution to predict Gmm and Gmb at initial, design and max gyrations. The work developed ANN models in which 18 inputs variables were composed of sieve passing (gradation of the mix), bulk specific gravity of aggregates, performance grade of the binder (low and high), binder content of the mix and the target number of gyrations (initial, design and max). This study used 1617 different mix design to develop an ANN with three hidden layers composed by (300-300-600) neurons. In addition, an independent validation of this ANN model was made with 200 mix designs that were set aside from data used for training. This ANN model showed a performance of R^2:

1. values from 0.816 to 0.841 for Gmb at $N_{initial}$, N_{design} and $N_{maximun}$
2. 0.758 for Gmm

However, after validation the R^2 of the model decreased to:

1. Values range from 0.6361 to 0.675 for Gmb at $N_{initial}$, N_{design} and $N_{maximun}$
2. 0.531 for Gmm

Also, the sensitivity analysis of the inputs variables showed that the most sensitive input was the gradation for the role during compaction, followed by Gsb, the PG, asphalt binder content (Pb) and Number of gyrations. Nevertheless, the reduction of R^2 could indicate that some grade of overfitting was found in the model and wrong conclusions could be achieved based on this model.

Furthermore, the work of (Sebaaly, Varma, & Maina, 2018) developed ANN models using 444 Marshall mix design data. These ANN models were optimized applying the Genetic Algorithm (GA) concept in which 12 inputs variables were considered like sieve passing (gradation of the mix), Gsb, the high-performance grade of the binder, and Pb to predict four outputs (air voids (Va), Marshall stability, Marshall Flow and Gmm). Additionally, despite the quantity of data pairs and the small structure of the ANN model developed which just one hidden layer was considered with 25 neurons, the model appears adequately performed when a validation process was applied to verify predictions with new test lab results.

However, it remains unclear if the procedure used to build the dataset. The work does not explain if some values were interpolated, obtained by running tests or estimated with common formulation like Gmm in the Marshall design method.

Moreover, recently the study of (Fadhil, Ahmed, & Al Mashhadany, 2022) built a ANN model to predict Gmb, air voids (Va), Marshall stability, Marshall flow and Marshall stiffness by using 5 inputs variables (Surface area, penetration of the binder, Kinematic viscosity, Pb and Mechanical abrasion). This model was developed using 170 asphalt mixes and an ANN structure based on one hidden layer of 30 neurons. The model shows a very high R^2 for training. However, no validation process with unseen data is shown to avoid drawing wrong conclusions based on an overfitted model.

1.5.2 Discussion of the application of AI concepts in the field of HMA

It is known that a small test set could produce statistical ambiguity regarding test error estimation, thus being unclear if one algorithm generalizes better than another for the same task when different algorithm architectures are evaluated (Goodfellow et al., 2016). For this reason, the amount of data necessary to correlate HMA performance with its components (aggregates, binder) properties should be based on hundreds of thousands of examples (observations-pairs) to avoid overfitting issues based on the suggestions in the field of computer science to develop AI models (Goodfellow et al., 2016). Moreover, researchers expose that this statistical ambiguity is not a serious issue when hundreds of thousands of examples (pairs) were used.

The HMA performance test results (data) are difficult to collect because they are complex and take a large amount of time and effort. In addition, because performance tests are expensive for HMA suppliers (asphalt plant), the tests like fatigue and dynamic modulus are not a common practice in HMA design or quality control, thus limiting the data available due to only tests are running for research projects. Thus, many research projects have tried to directly correlate HMA volumetric parameters with performances (fatigue and rutting) under

traditional statistics methods. However, these correlations remain unclear, thereby designing HMA to achieve good performance becomes difficult.

Moreover, researchers have been developing ANN models to predict volumetric properties Gmm and Gmb. Nevertheless, these models have been developed by combining two different phenomena in the same ANN model. Theses volumetric properties (Gmm and Gmb) are determined with different tests because separate phenomena are implied. The former Gmm is obtained following the Rice test in which no compaction is done, the asphalt mix is separated into loose, individually bonded aggregates while Gmb is obtained by compacting the asphalt mix with a specific energy. For this reason, it would be appropriate to develop ANN models separately to predict Gmm and Gmb.

Furthermore, it seems that in the field of HMA, significant limitations in the application of basic MLTs procedures, have allowed researchers drawing wrong conclusions based on overfitted models in which the ability to produce a small error in the training and reducing the gap between training and test error is not validated since the MLTs basic procedures like 1) displaying error curves, and 2) validating with examples from other projects (unseen data by the model), apparently were underestimated by researchers in the study of HMA phenomena under AI concepts. It was seen in the literature that merely evaluating the performance of an algorithm only based on the correlation coefficient R^2 as a performance indicator and the root mean square error (RMSE), is not enough. Thus, these performance indicators could show perfect values as MRSE close to 0 and R^2 equal to 1 if they are measured in an overfitted model.

Researchers in the field of computer science and engineering have found several ways to solve this problem about fewer dataset by applying:

1) The basic procedure of the cross-validation,
2) The MLT referred to random forests as (Gong et al., 2019) suggested when to explore the impact of asphalt mixtures properties on pavement performance; this MLT, the Random Forest, is a classification algorithm which involves a collection of tree-structured classifiers; these trees are independent and distributed identically random vectors, thus providing a unit vote for the class of input vectors (arrays) by each tree (Yeh, 2007),

3) The transfer learning concept in which high-performance algorithms are trained with more easily collected data from different domains (Weiss, Khoshgoftaar, & Wang, 2016). In fact, for ANN the transfer learning concept is focused on storing knowledge from the source phenomenon, in form of tuning weights of the different ANN-layers by transferring these weights to the target phenomenon. Thus, a small number of new data from the target phenomenon is required to train the pre-trained weights due to the common practice of randomly initialized weights is avoided, and an optimization of the training process is achieved. (Mo et al., 2018). This concept avoided developed models ANN models from scratch.

1.6 Summary

The HMA design has two approaches, volumetric and balanced mix design. The former, the volumetric HMA design which has been optimized considering the aggregate packing in dry condition by avoiding the role of asphalt binder, while the latter the Balanced Mix Design (BMD) deals with running expensive performance tests to validate the state of voids. Many researchers have tried to apply MLTs to develop models that could predict HMA performance with the idea of optimizing the HMA design by avoiding expensive tests. However, These MLTs have not been properly applied following the general process of machine learning in materials science (Sample construction, Model building, Model evaluation).

Therefore, developing models to predict HMA performance using AI needs a good understanding of some concepts like: 1) the cross-validation concept which should be essentially applied when a small test set is used to avoid producing statistical ambiguity in the average test error estimation, 2) the learning curves by assessing models 'performance to avoid overfitting or underfitting problems by understanding how the model architecture and quantity of data influence on these curves' concepts and 3) the validation of the model developed by the extent to which the model can achieve a good performance on the real unseen data. These concepts were presented and studied to evaluate how MLTs could adequately apply to optimize HMA design.

The one of the benefits of this research is to understand that how the MLT specially ANN can be applied in the field of HMA to adequately predict outputs that can be useful to optimize the HMA design. The result of evaluation of this complex tool (ANN) in HMA will help to reach the modification to be able to adequately predict behaviours like rutting, fatigue, permeability, workability and resilient modulus.

In the following chapter, a description of the sequential methodology is provided to set up the Gmm and the Gmb prediction models. This chapter 2 summarizes the steps associated with sample construction, model building, and model evaluation. These two outputs (Gmm and Gmb) were selected due to the quantity of data found (hundreds of thousands of examples (pairs)) which allows the author to properly apply the AI concepts studied.

1.7 Objective of the study

A vast database regarding laboratory tests is registered in the LTPP database (FHWA, 2019), efficient techniques that can adequately predict complex phenomena are currently accessible and also since the test methods for obtaining Gmb and Gmm have been normalized by AASHTO and ASTM in which variation in the test conditions has been reduced, the main objective of this work is to optimize the volumetric HMA design by separately predicting the bulk specific gravity (Gmb) and the maximum specific gravity (Gmm).

These two volumetric properties are still obtained running test lab due to limitations found regarding models developed by researchers like:

1. Low performance (R^2) of models developed by (Ozturk & Kutay, 2014) in which models probably exceed the multilaboratory precision values suggested by (FHWA, 2010) for Gmm and Gmb
2. unclear methodology to build the database (Sebaaly et al., 2018). In addition, two phenomena were mixed (Marshall compaction and Rice test) into the ANN model developed
3. Small quantity of data to apply ANN method to predict the Gmb for Marshall design method (Fadhil et al., 2022) by which only 170 asphalt mixes were used and lack of validation with unseen data was displayed.

Two models ANN were developed for this purpose by evaluating two different phenomena separately (compaction-SGC and no-compaction-Rice test). These two models consider the aggregate and binder properties as independent variables; also, the maximum specific gravity (Gmm) and bulk specific gravity of the HMA (Gmb) as dependent variables were predicted by these two models separately. In addition, the volume of data utilized in this study to overcome limitations of ANN method, is almost:

1. 195 times higher than quantity of data employed by (Ozturk & Kutay, 2014) to predict Gmb and 174 times higher to predict Gmm
2. 713 times higher than volume of data used by (Sebaaly et al., 2018) to predict Gmm
3. 156 times higher than volume of data utilized by (Fadhil et al., 2022) to predict Gmb

This DNN method permits observing which material properties have more influence into the HMA volumetric properties (Gmb and Gmm). These two values are essentials to obtain the state of voids in the HMA design. Thus, achieving this objective the study will be a very important step towards the goal of using AI in order to reduce the lab work and make the process more efficient by virtually automating crucial parts of the process in the traditional trial error HMA.

The ANN approach was used in this study because it was sufficient data to train, and predicting Gmb and Gmm, is a nonlinear problem. Also, ANNs provide an analytical alternative to statistical methods which require higher background of statistics with strict assumptions of normality, linearity, variable independence. Moreover, the ANN has the capability of capturing nonlinear and complex relationship between dependent and independent variables of any physical process with accuracy by allowing the user to model phenomena that could be extremely complicated to explain by traditional statistical methods (Maind & Wankar, 2014). Furthermore, it was found in the literature that ANN method has been applied by researchers in the field of HMA probably because this algorithm does not require any a priori model or any model structure before starting the ANN model. However, ANN is limited to no ability to extrapolate results and large datasets are essential to perform well.

CHAPTER 2

METHODOLOGY

Based on the literature, this research project aims to evaluate the use of artificial intelligence (AI), specifically deep neural network (DNN) to predict the HMA volumetric properties Gmm and Gmb based on the data available in the field of HMA. The DNN, as a supervised machine learning algorithm, can successfully generalize the phenomenon when large amounts of data can support it. For this reason, to address the limitation observed in the literature, thereby avoiding overfitting models, the author considers applying the model evaluation concepts proposed by Zhao et al. (2017) as mentioned in the section 1.4.2.1 by:

1) displaying an error curve evaluation of all models developed; and
2) validating the model selected as optimum performance using data from other projects (unseen).

The cross-validation procedure was not considered because the data set used throughout this study contained thousands of examples (pairs) to develop the models. This process is exposed in this chapter and its results are displayed in chapter 3.

To address the limitations associated with packing evaluation under dry condition, data quantity and overfitting observed in chapter 1, a research methodology has been established (Figure 2.1). The work consists in, first, the development of two deep neural networks (DNN) for the volumetric prediction of Gmb and Gmm, and then the selected models are validated. In order to develop the working DNNs, two datasets were built to overcome the constraints observed regarding the amount of data. Additionally, one dataset unseen by the models was built and used to remedy the lack of external validity when wrong conclusions were probably drawn by assessing neural networks overfitted in the field of HMA. Those datasets contain bulk and max specific gravity test results which were labelled with the corresponding aggregates and binder properties. The two DNNs models were developed applying a supervised learning approach, described in chapter 1, in an effort to optimize the volumetric HMA design considering the influence of asphalt binder content into the HMA aggregate packing phenomenon, instead of only optimize the aggregate packing (dry condition) as found

in previous studies. The framework displays the model development of: 1) SGC test DNN-model to predict Gmb; and 2) Rice test DNN-model to predict Gmm, by following the three phases. In phase 1, the dataset (inputs-outputs) in which volumetric properties of HMA were labelled with the properties of materials components was constructed. In phase 2, empirical experiments were run by changing the neural network architecture, thus achieving an optimum model for each prediction (Gmb and Gmm). In the third phase, a validation process was established in which models were tested using unknown data to avoid overfitting problems. Finally, a correlation analysis was suggested for models validated to evaluate the influence of the input variables on the outputs (Gmm and Gmb) which are the main HMA volumetric properties indispensable to establish the state of voids in the HMA volumetric design.

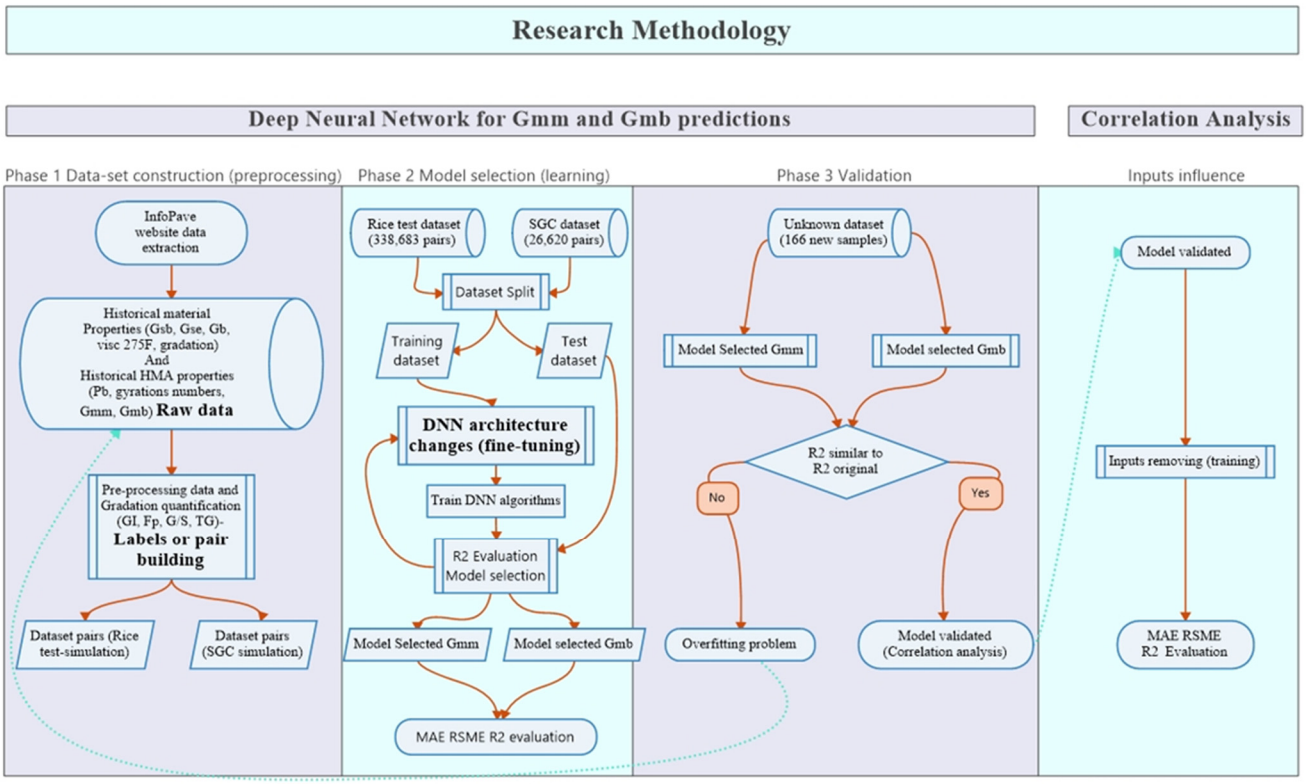


Figure 2. 1 Methodology framework in which three phases are presented and a correlation analysis process for the model selected

2.1 Dataset construction (phase 1)

The two datasets used throughout this study to develop the Gmb and Gmm prediction models and the validation dataset are based on structured data in which the features were extracted manually. This was done by applying shallow machine learning methods due to models learnt from data described by predefined features in which manual human intervention to collect the features is necessary because this process requires previous knowledge of the data. This Shallow Learning differs from deep learning due to the latter method extracts the features using computed algorithm without any manual human intervention; in other words, a deep learning algorithm will automatically recognize the weighted features.

2.1.1 Historical database for Gmb and Gmm prediction models

The historical database for both prediction models (Gmb and Gmm) was built extracting raw data directly (properties of HMA, aggregates and binder) from the InfoPave website which is supported by The Federal Highway Administration into The Long-Term Pavement Performance (LTPP) program. This was done with the SQL export feature in which data tables containing the HMA material information could be linked with the volumetric test result (Gmb and Gmm) as can be seen in Figure 2.2. The columns of the Figure 2.2 represent HMA and component properties as well as test parameters (1-Gyration number, 2-Effectve asphalt, 3-VMA, 4-Gmb, 5-Pb, 6-Va, 7-Gse, 8-Gsb, 9- 1” ½ passing sieve , 10-1”passing sieve, 12- 3/4” passing sieve, 13- ½” passing sieve, 14- 3/8” passing sieve, 15- #4 passing sieve, 16- #40 passing sieve, 17- #10 passing sieve, 18- #80 passing sieve, 19- #200 passing sieve, 20-viscosity 275 F, 21-Gb.

Moreover, because one of the objectives of this work is to optimize the volumetric HMA design considering the influence of asphalt binder content into the HMA aggregate packing phenomenon, the raw data was collected by limiting confounding variables into the phenomena correlations analysis based on:

1) HMA designs based on virgin aggregates only; designs using Reclaimed Asphalt Pavement (RAP) were not considered in these databases,

2) HMA design based on continuous and discontinuous gradations were considered,

3) Only HMA designed under Superpave method were selected to simulate SGC,

4) Only aggregate volumetric properties were considered. In effect, no aggregate surface texture or angularity were considered.

The data were preprocessed manually to minimize problems, noise problems regarding the quality of the data collected. This preprocessing consisted in cleaning the data and missing values were completed.

SUPER_NDES	EFF_ASPHALT_CONT	VOIDS_MINERAL_AGGR(ASTM)_SPEC_GRAVITY	BULK_SPEC_GRAVITY	ASPHALT_CONT	PCT_AIR_VOIDS	EFFECT_SPEC_GRAVITY	AG_COMB_BULK_SPEC	ONE_AND_HALF_PASSING	ONE_PASSING	THREE_FOURTHS_PASSING	ONE_HALF_PASSING	THREE_EIGHTHS_PASSING	NO_4_PASSING	NO_40_PASSING	NO_10_PASSING	NO_80_PASSING	NO_200_PASSING	NO_10_PASSING	VISCOSITY 275	ASPHALT SPECIFIC GRAVITY
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	97	91	83	49	16	34	8	4.9000001	34		1.029999971
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	93	85	76	45	16	33	8	4.9000001	33		1.029999971
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	99	95	90	48	15	29	8	6.0999999	29		1.029999971
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	95	86	80	52	18	36	10	5.9000001	36		1.029999971
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	100	94	85	55	19	38	11	6.9000001	38		1.029999971
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	94	88	83	50	15	30	9	5.80000019	30		1.029999971
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	95	92	87	50	16	31	9	5.5	31		1.029999971
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	92	86	81	50	18	35	11	6.5	35		1.029999971
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	99	95	88	56	20	40	11	5.80000019	40		1.029999971
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	95	89	82	50	18	34	9	5.4000001	34		1.029999971
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	98	91	84	51	17	34	9	5.4000001	34		1.029999971
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	97	93	87	56	18	37	9	5	37		1.029999971
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	96	91	85	55	19	34	10	4.9000001	34		1.029999971
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	94	82	73	42	14	26	8	5.0999999	26		1.029999971
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	97	90	80	45	14	29	8	5	29		1.029999971
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	97	91	83	49	16	34	8	4.9000001	34		1.029999971
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	93	85	76	45	16	33	8	4.9000001	33		1.029999971
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	99	95	90	48	15	29	8	6.0999999	29		1.029999971
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	95	86	80	52	18	36	10	5.9000001	36		1.029999971
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	100	94	85	55	19	38	11	6.9000001	38		1.029999971
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	94	88	83	50	15	30	9	5.80000019	30		1.029999971
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	95	92	87	50	16	31	9	5.5	31		1.029999971
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	92	86	81	50	18	35	11	6.5	35		1.029999971
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	99	95	88	56	20	40	11	5.80000019	40		1.029999971
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	95	89	82	50	18	34	9	5.4000001	34		1.029999971
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	98	91	84	51	17	34	9	5.4000001	34		1.029999971
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	97	93	87	56	18	37	9	5	37		1.029999971
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	96	91	85	55	19	34	10	4.9000001	34		1.029999971
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	94	82	73	42	14	26	8	5.0999999	26		1.029999971
109	4.300000191	14.19999981	2.69499993	2.584000111	4.5	4.1	2.915999889	2.911999941	100	100	97	90	80	45	14	29	8	5	29		1.029999971
126	4.699999809	14.89999962	2.47099996		5	4.1	2.671000004	2.648000002	100	100	90	69	60	41	11	26	7	4.5	26	623.90002	1.026000023
126	4.699999809	14.89999962	2.47099996		5	4.1	2.710999966	2.572999954	100	100	90	69	60	41	11	26	7	4.5	26	623.90002	1.026000023
126	4.699999809	14.89999962	2.47099996		5	4.1	2.671000004	2.648000002	100	100	91	70	58	40	10	25	7	4.4000001	25	623.90002	1.026000023
126	4.699999809	14.89999962	2.47099996		5	4.1	2.710999966	2.572999954	100	100	91	70	58	40	10	25	7	4.4000001	25	623.90002	1.026000023
126	4.699999809	14.89999962	2.47099996		5	4.1	2.671000004	2.648000002	100	100	90	75	67	49	13	29	9	6.4000001	29	623.90002	1.026000023
126	4.699999809	14.89999962	2.47099996		5	4.1	2.710999966	2.572999954	100	100	90	75	67	49	13	29	9	6.4000001	29	623.90002	1.026000023

Figure 2. 2 Illustration of raw data example for SGC simulation, collected from the InfoPave website supported by The Federal Highway Administration into The Long-Term Pavement Performance (LTPP) program

1) Data cleaning:
   a) Eliminating duplications or similar pairs. These pairs were the pairs which had the same inputs and outputs. The pairs which had at least one input variable or output variable different was stored;
   b) Incoherent values regarding HMA volumetric properties (Gmb greater than Gmm) were removed;
   c) Illogical values regarding aggregates volumetric properties (Gsb greater than Gse, Gsb greater than Gsa or Gse grater than Gsa) were eliminated;
   d) Incoherent values regarding Gmb and Gmm and the states of voids (VMA, Va and VFA).

2) Missing values:
   a) When Gmm was not found, but Gmb and the stated of voids were found, Gmm was calculated to create the pair;
   b) When Gmb was not found but Gmm and the stated of voids were found, Gmb was calculated to create the pair;
   c) Complete rows of data were eliminated when Gmm and Gmb were missing;
   d) When Gsb was not found but Gse, Gb and the asphalt absorption (Pba) were available, Gsb was calculated to complete the pair;
   e) When Gse was not found but Gsb, Gb and the asphalt absorption (Pba) were available, Gse was calculated to complete the pair;
   f) Complete rows of data were eliminated when Gse and Gsb were missing.
   g) When Gb was not found but Gmm, Gse, and Pb were available, Gb was calculated to complete the pair;
   h) When Gb was not found but Gse, Gsb and the asphalt absorption (Pba) were available, Gb was calculated to complete the pair;
   i) When Pb was not found but Gmm, Gse, and Gb were available, Pb was calculated to complete the pair;
   j) When viscosity was not found, the rows were eliminated;

k) No values were estimated. Thereby, neither interpolation method nor mean, median or mode value of the respective feature were applied to fill missing values to complete pairs.

Missing values were calculated following the equations presented in the Asphalt Institute manual SP-2 (Asphalt-Institute, 2001).

In effect, asphalt binder, HMA and aggregate properties like Gse, Gsb, were collected and directly stored. However, the aggregate gradation collected was processed to obtain the quantification factors such as TG, GI, Fp and G/S and then stored. This secondary data was calculated based on the literature review regarding gradation quantification which allowed quantify the aggregate blend as a number by which correlation between performance and gradation were found.

This aggregate gradation operationalization approach (quantification) was realized employing the gradation index (GI), the gravel-sand (G/S), the Ramcodes characteristic factor (Fp) and the total gradation (TG) concepts.

1) Gradation index (GI) which is a ratio of the area under the gradation curve and the total area

$$GI = \frac{a}{A} * 100 \qquad (2.\ 1)$$

Equation 2. 1 Gradation index calculation taken from (Setiawan et al., 2017)

Where:

A is the total area (mm^2),

a) "a**"** is the area retained of the gradation curve (mm^2) as can be seen in Figure 2.3 by illustrating the shaded area above the gradation curve; it is calculated according to the equation 1.4 mentioned in the section 1.2.3 of the chapter 1 (Gradation Index concept).

$$a = \sum_{i=0}^{n} \left(\frac{Sr_i + Sr_{i+1}}{2}\right)(T_i - T_{i+1}) \qquad (1.4)$$

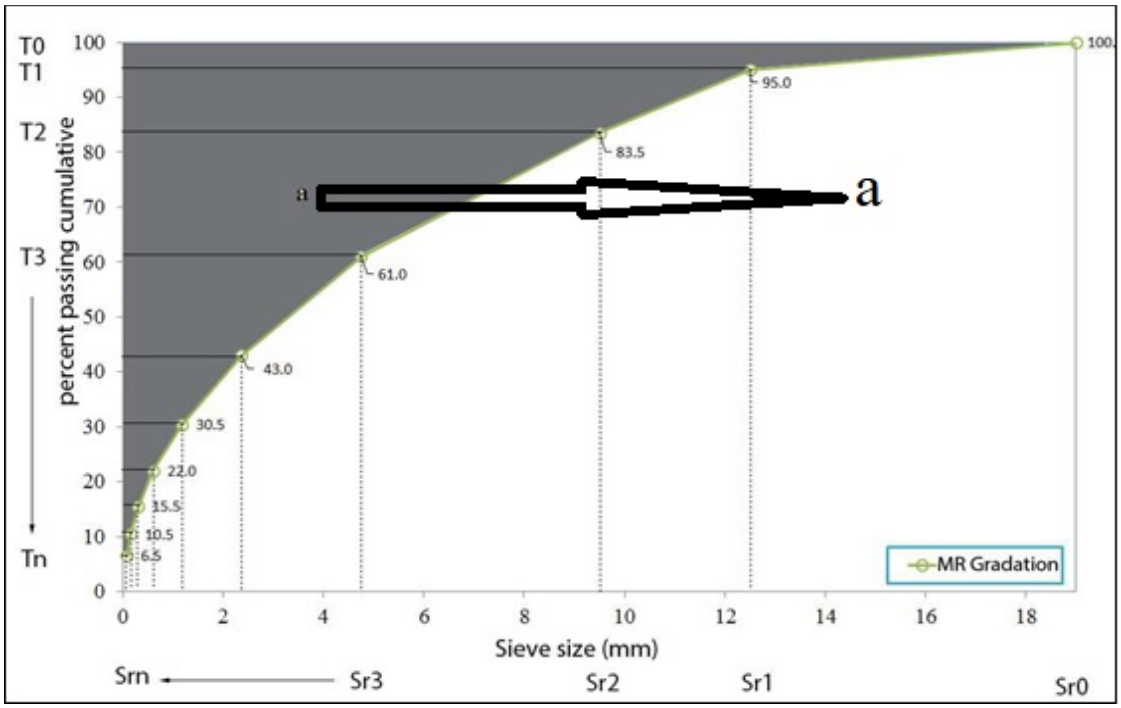


Figure 2. 3 Illustration of the retained area of gradation curve taken from (Setiawan et al., 2017)

b) Sr is sieve size (mm),

c) T is the cumulative retained aggregate in which 10% is equal to 10mm (mm).

d) MR gradation or mid-range is the curve example which is based on the 2010 Indonesian Highway Specification for asphalt concrete wearing course *(Setiawan et al.,* 2017*)*

2) Grave-Sand ratio (*G/S*) which the ratio of Gravel content divide by sand content; being:

a) Gravel content (*G*) :

$$G = 100 - passing\ sieve\ 4.75 \quad (2.\ 2)$$

b) The finer content (*F*):

$$F = passing\ sieve\ 0.075 \quad (2.\ 3)$$

c) The sand content (*S*):

$$S = passing\ sieve\ 4.75 - F \quad (2.\ 4)$$

$$\frac{G}{S} = \frac{100 - passing\ sieve\ 4.75}{passing\ sieve\ 4.75 - F} \quad (2.5)$$

Equations 2.2, 2.3, 2.4 and 2.5 Taken from (Sánchez-Leal 2007).

3) Ramcodes characteristic factor (*Fp*) which is a linear combination of gradation, represented as:

$$Fp = 1.1 * \frac{F}{(1+G)} \quad (2.6$$

Where:

a) *G* the Gravel content from equation 2.2
b) *F* the finer content from equation 2.3

Equation 2.6 Taken from (Sánchez-Leal 2002).

4) Total gradation (*TG*) which the sum of all sieve passing from 20mm to 80 microns

$$TG = \sum passignform\ sieve\ 20mm\ to\ 80\mu m \quad (2.7)$$

Equation 2.7 Taken from LC 26-007 Test Method (MTQ 2019).

2.1.2 Validation dataset (unseen by models selected

A dataset of 166 pairs was used to validate whether the models selected were overfitted or not; it was built following the same preprocessing data regarding the aggregate gradation operationalization approach (quantification) to obtain the four gradation factors (GI, FP, G/S and TG) as shown previously. However, the data regarding material properties were collected directly from suppliers' reports and mix design labs reports, thereby avoiding the preprocessing data to minimize noise problems (Data cleaning and Missing values) found previously in data collected from infoPave. This unseen dataset was originally produced in Mexico and Quebec for HMA design used in projects.

2.2 Model development (phase 2)

The historical database for Gmm prediction models contains 316,594 pairs and the database for Gmb contains 26,620 pairs. Both datasets were split in the training and testing data. The dataset of 316,594 pairs regarding Gmm model was split by dividing the dataset in 60 % and 70 % for training and the residual data 40%-30% for the test. As for the dataset of 26,620 pairs concerning SGC model, it was split into 75%-70% for training and 25-30% for the test. The two models were developed under trial-and-error approach to find the number of nodes and the number of hidden layers, in search of the optimum model. Moreover, a fine-tuning hyperparameter regarding the dropout was realized, thus examining the strengths and weaknesses of the developed models through R2 in which the predicted values were compared with the actual values and indicating that a higher R2 value was considered a better fit of the development data set. Figure 2.4 shows the process framework to select the optimum model.

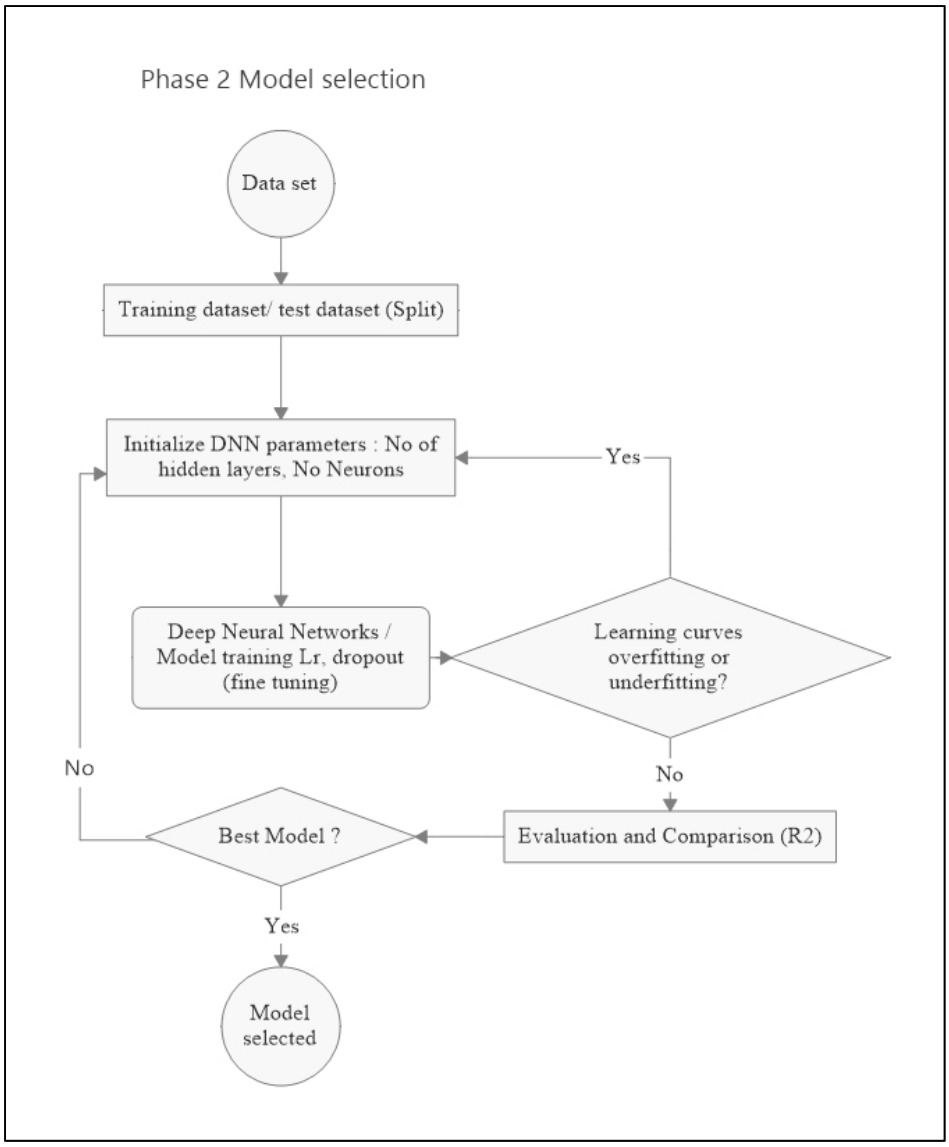


Figure 2. 4 The flow diagram to select the optimum model

In addition, the input variables were selected based on the quantity of data available in infoPave that could help to avoid models overfitted when a small quantity of data is used to generalize a complex phenomenon. Moreover, this data input was collected with the main objective of determining correlations between HMA materials components (aggregates and binder) properties and (Gmm and Gmb) as output variables, where input variables concepts were:

a) Aggregates
   - ✓ Gradation

- ✓ Volumetric properties

b) Binder

- ✓ Content
- ✓ Volumetric properties
- ✓ Viscosity. This variable considers the influence of the temperature on the Gmb and Gmm predictions

Applying this principle, the following input variables were selected for each model developed:

1) Gmb prediction

a) Inputs

- ✓ Number of gyrations (*N_G*). This variable is important because it exists a specific Gmb produced by each gyration of the SGC equipment. This SGC is a compaction phenomenon by which a HMA increases its density by applying 600-kPa of pressure at 30 gyrations per minute. This SGC test has been standardized by following ASTM D6925 and AASHTO T312.
- ✓ Aggregate properties
  - Bulk specific gravity for the total aggregate (*Gsb*). This variable gives the neural network the volumetric property of aggregate blend of the HMA.
  - Aggregate gradation through the secondary data (*GI, G/S, Fp, TG*). These variables quantify the aggregate skeleton of the HMA.
- ✓ Binder
  - Asphalt content (*Pb*)
  - Viscosity at 275 ºF , this variable informs the neural network regarding binder behaviour at temperature mentioned
  - Specific gravity of asphalt (*Gb*), this variable gives the neural network the volumetric property of asphalt binder into the HMA

b) Outputs

- ✓ Bulk specific gravity of the HMA (*Gmb*)

2) Gmm prediction.
    a) Inputs. The role of this inputs is the same as explained for Gmm model except for the Number of Gyration (*N_G*) which is not part of the Rice test to obtain Gmm. This Rice test has been standardized by following ASTM D2041 and AASHTO T209 (Asphalt-Institute, 2014). The Rice test is based on a vacuum process by which mix asphalt is individually separated into loose coated aggregate, thus displaying that this test is not physically a HMA compaction phenomenon to achieve a HMA with no air void inside.
        ✓ Aggregate properties
            - Bulk specific gravity for the total aggregate (*Gsb*)
            - Aggregate gradation through the secondary data (*GI, G/S, Fp, TG*)
        ✓ Binder
            - Asphalt content (*Pb*)
            - Viscosity at 275 ^{o}F
            - Specific gravity of asphalt (*Gb*)
    b) Outputs
        ✓ Maximum specific gravity of the HMA which is no air voids inside (Gmm)
        ✓ Effective specific gravity of aggregate (*Gse*)

The initial neural network architecture implemented in this study was suggested based on the common practice in the field of machine learning because of the amount of data pairs collected permitted having a well-early approach to generalize the phenomenon. Therefore, four (4) hidden layers into a neural network with more than 150 neurons per hidden layer were built in which rectified linear unit (ReLU)was used as activation function to train the model because ReLu function does not activate all the neurons at the same time when negative values are included into the datasets. Thereby, neurons do not get activated for negatives inputs values. This characteristic of ReLu function makes more efficient the ANN-model compared to others common function like the sigmoid and tanh function (Brownlee, 2019) when no negatives values are used as inputs to training models. This initial process was done not according to any rule, but because they gave reasonable results regarding the learning curves in which the loss curve of the model was almost always lower on the training (blue) than the test (orange)

dataset, thereby indicating a small overfitting as can be seen in Figure 2.5. The loss function to calculate the model error in this research was MSE which is the default function loss to use for regression problems. In addition, mathematically, it is the preferred loss function when the distribution of the target variable is Gaussian (Brownlee, 2019). The number of epoch defines the number of complete passes that the learning algorithm runs through the entire training dataset (Brownlee, 2019). The Figure 2.6 illustrates the learning curve concept based on the loss function to know how evaluate the initial hyperparameters learning rate (LR) and the size of the batch. Thus, this study empirically selected the LR= 0.00010 and a batch equal to 10 by comparing the initial curve of the neural network architecture implemented as can be seen in Figure 2.5 with the learning rate concept as shown in Figure 2.6.

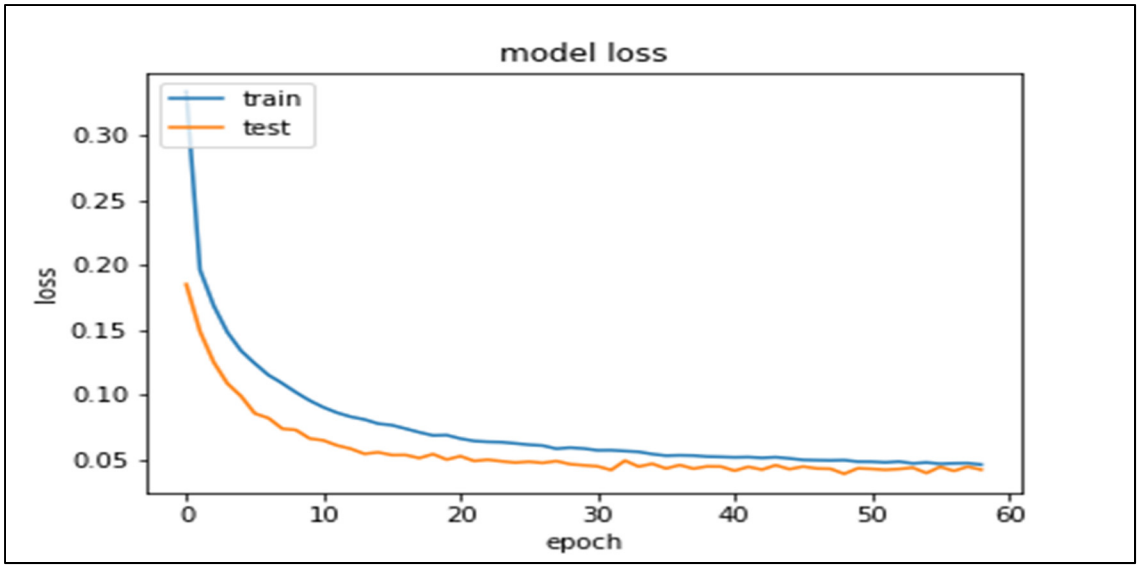


Figure 2. 5 Illustration of an early approach by using four (4) hidden layers into a neural network with more than 150 neurons per hidden layer for Gmb prediction model.

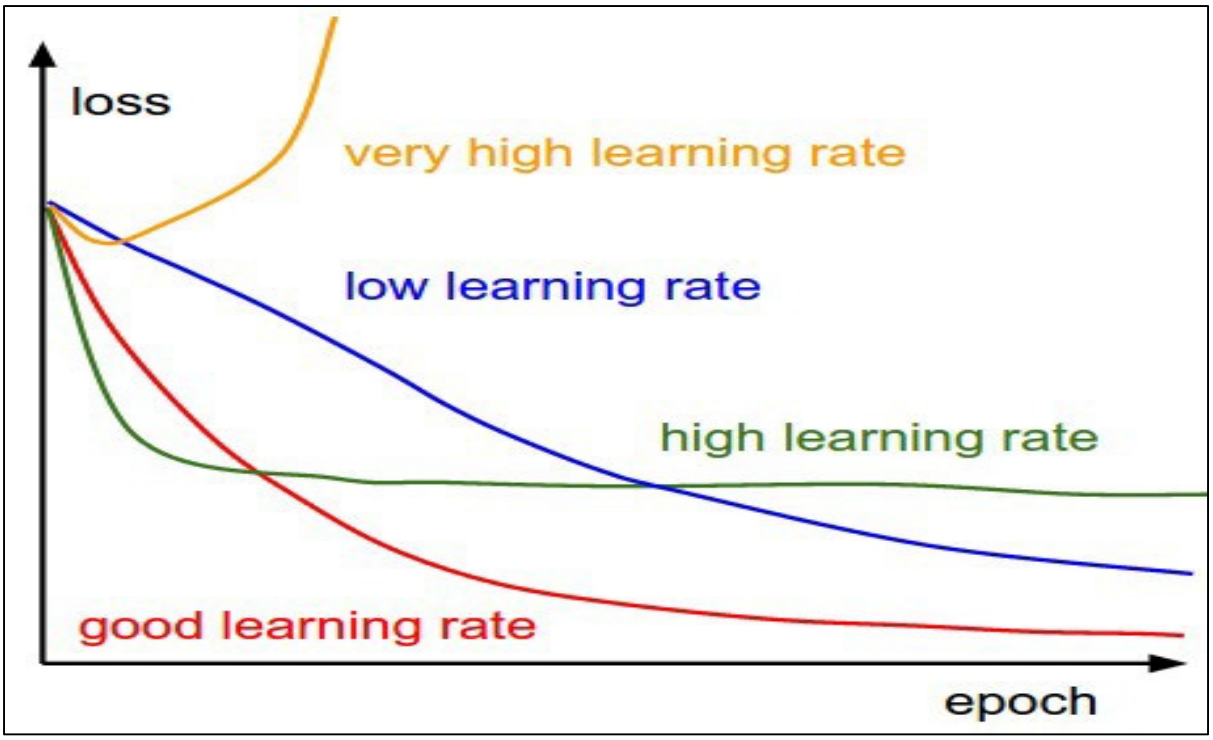


Figure 2. 6 Illustration of the effect of learning rate on the loss function taken from (Stanford-University, 2021)

Researchers have been applied ANN concepts in the field of HMA. For instance, the study of (Zavrtanik et al., 2016) developed models in which the neural networks had between two and three hidden layers with five to 20 neurons per hidden layer by using data pairs between 176 and 1634 data pairs. The hyperparameter learning rate (LR) used for training was between 0.1 and 0.5. However, it remains unclear the selection of these values adopted for tunning the models because any learning curve was displayed to evaluate the model performance generalization. Furthermore, the work of Al-Mosawe (2016) created a model with three hidden layers and five to 15 neurons per hidden layer in which a maximum of 22 data pairs were used from the HMA studied. Also, this work did not show a selection method either the leaning rate (LR) nor the batch.

Apparently, these studies showed they do not allow developing architectures with more neurons due to the risk of overfitting problems. However, overfitting issues were not evaluated by these two studies.

Although, the work of (Ceylan et al., 2007) displayed learning curves to evaluated performance generalizations regarding HMA Dynamic Modulus Prediction Models, the study did not show either the initialization method for hyperparameters nor the value of these hyperparameters (LR and batch) used in the ANN developed.

In summary, the constrains observed in the literature when AI methods were applied into the field of HMA were overcome by 1) using a large quantity of data in this research, 2) initializing hyperparameters based on (Stanford-University, 2021) approach regarding the effect of learning rate on the loss function and 3) evaluating performance model generalizations through learning curves.

2.2.1 Program and toolkit

Models were trained using TensorFlow software by applying the deep learning framework of Keras under Python libraries programing code. This allows working on the two phenomenon's problems (Rice test and HMA compaction process) with multiple inputs variables in which a set of input values is processed together to obtain an output value. This concept is referred to matrix operations in the machine learning field by which six models were developed to predict Gmb and eight models were developed to predict Gmm values.

2.2.2 SGC model development

SGC model was developed by empirically training six (6) different models as it is shown in Table 2.1 to evaluate performance in which:

1) Neural architecture as shown in Figure 2.8 was changed in:
   a) number of hidden layers (by 4-5-6 layers),
   b) number of neurons (by 150-200-350-450 nodes),
2) data split (75% or 70% of data for training and 25%-30% data for test),
3) fine-tuning by changing the hyperparameter dropout (0.10-0.15).

The Dropout is a hyperparameter which prevents overfitting and provides an alternative to combine different neural network architectures efficiently. The process of dropping out units (hidden and visible) in a neural network is referred to “dropout”. This process of dropping a unit out during training consists in removing it from the neural network, as well as all its connections (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014), as shown in Figure 2.7. In addition, the extent to which neurons (units) are selected for dropping is random (Srivastava et al., 2014). This regularization method was applied to avoid overfitting problems since the application of dropout concept regularization for reducing overfitting and improving the generalization of deep neural networks is a common practice in the field of computer science (Brownlee, 2019). An illustration of the architecture neural network of model 6 can be seen in Figure 2.8.

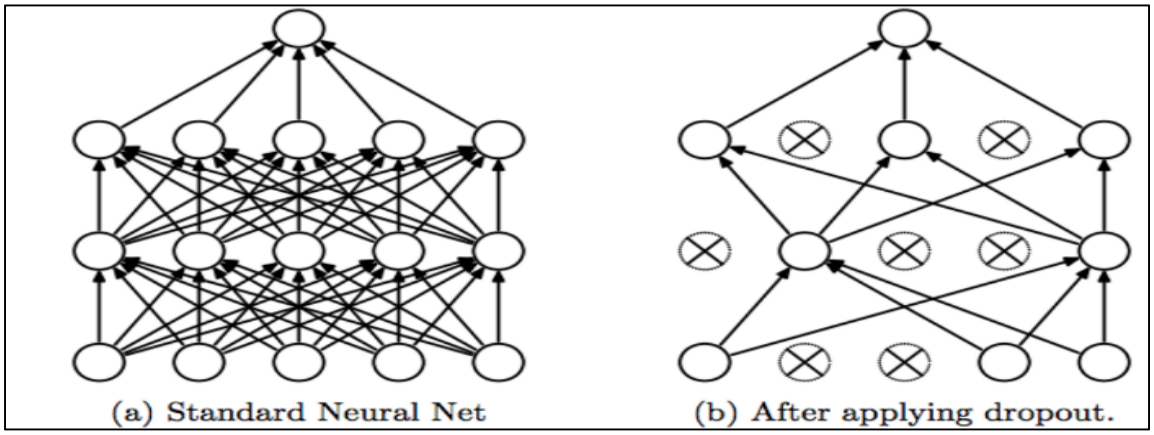


Figure 2. 7 Dropout Neural Net Model. Left: A standard neural net with 2 hidden layers. Right: An example of a thinned net produced by applying dropout to the network on the left. Crossed units have been dropped taken from (Srivastava et al., 2014).

Table 2. 1 Neural network changes for SGC Model developed

Model	Output	% training	Layers	Neurons	Dropout
1	*Gmb*	75	4	150	0.15
2			4	200	
3			5	350	
4		70	5	450	
5			6	450	
6			5	350	0.10

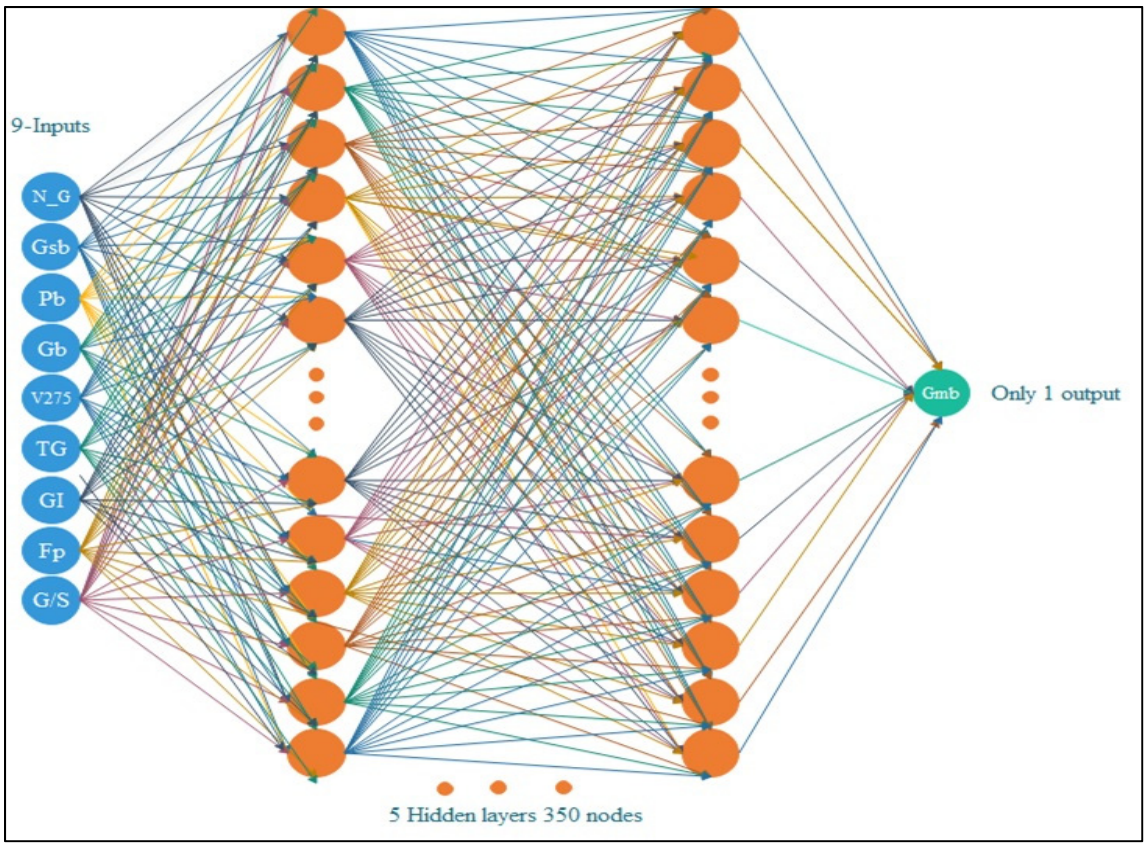


Figure 2. 8 Model 6 trained with 70% of data and tested with 30%, five hidden layers containing 350 neurons, Lr 0.00010, batch 10, dropout 0.10.

2.2.3 Rice test model development

Rice test model was developed by empirically training eight (8) models as can be seen in Table 2.2 to evaluate performance in which:

1) Neural architecture was changed in:
   a) number of hidden layers (3 layers or 2 layers),
   b) number of neurons per hidden layer (150 nodes and 250 nodes),
2) data split (70% or 60% of data for training and 30%-40% data for tests),
3) fine-tuning by changing the hyperparameter dropout (0.15-0.20),
4) output variables were changed (*Gmm or Gmm-Gse*)

Table 2. 2 Neural network changes for Rice test Model development

Model	Output	% training	Layers	Neurons	Dropout
1	*Gmm-Gse*	70	3	150	0.15
2	Gmm				
3			2		
4		60			
5		70		250	
6				150	
7		60			0.20
8	*Gmm-Gse*	70			0.15

An illustration of the neural network architecture for model 3, 4, 5, 6 and 7 can be seen in Figure 2.9; and the neural network illustration of model 8 is shown in Figure 2.10 .

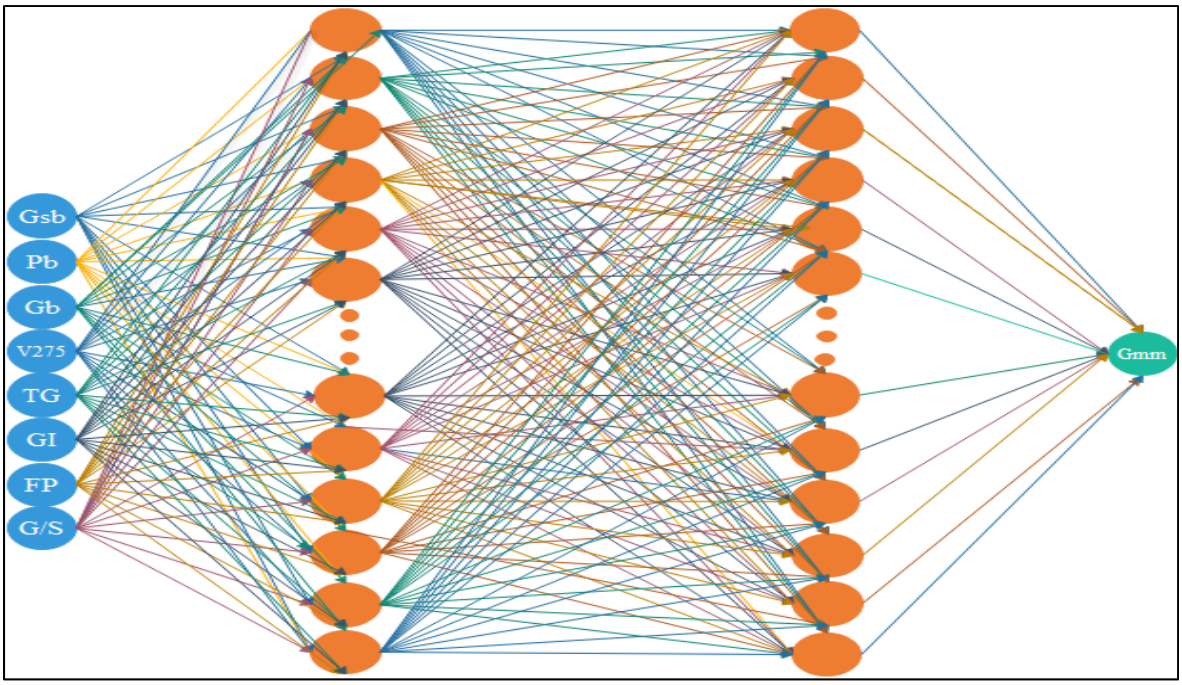


Figure 2. 9 Neural network architecture for **models 3, 4, 5, 6 and 7** in which only Gmm was applied as output

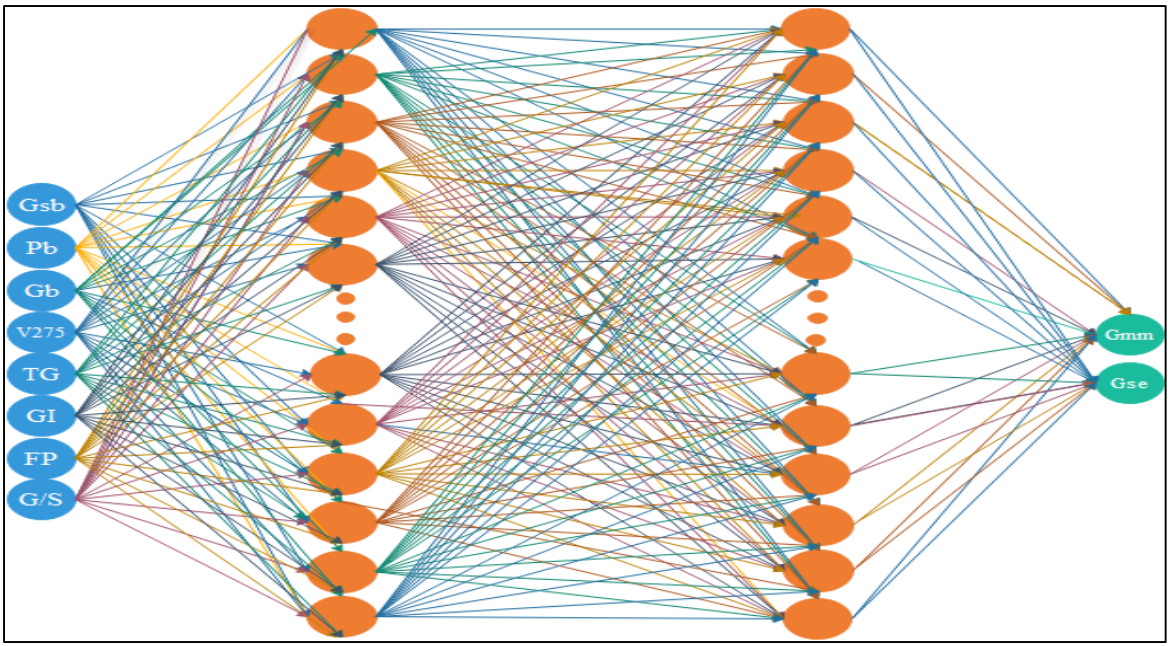


Figure 2. 10 Neural network architecture of **model 8** in which Gmm and Gse were applied as outputs.

2.3 Quality of the models (performance metrics)

Three performance metrics were adopted to adequately evaluate the accuracy of the regression models developed, thereby facilitating the process to elucidate the best performance model when changes in the neural network architecture were performed. These metrics were the coefficient of correlation of R^2, the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE). The two errors metrics were selected because both show an average model prediction error in units of the variable of interest, scale from 0 to ∞, and no indication of underperformance or overperformance of the model (the direction of errors) is presented by displaying that lower values are better. However, no comparison was made between RMSE and MAE due to the error increases quadratically in RMSE and contributes proportionally to the total error in MAE. This means that large differences between real and predicted values are penalized more in RMSE than in MAE. Thus, the optimum models for both prediction (Gmm and Gmb) were selected based on highest R^2 and the lowest values of RMSE and MAE.

2.3.1 The Coefficient of Correlation of R2

This coefficient was the main performance metrics in which predicted values were compared with original values from 0 to 1 interpreted as percentages. Thus, the closer to 1, the better the model performs. This coefficient concept was applied to compared models developed in which neural network architecture was empirically changed, thereby selecting the model with the highest R^2.

$$R^2 = 1 - \frac{\sum(yi-\hat{y})^2}{\sum(yi-\gamma)^2} \qquad (2.\ 1)$$

Where,

1) yi = actual value
2) $\hat{y}$ = predicted value of yi
3) γ= mean value of yi

2.3.2 The Root Mean Square Error (RMSE)

The Root Mean Square Error (RMSE) was calculated to consider the outliers (extreme values) into the model due to it is known that a constant value of Gmb, changes in Gmm by +0.01 would produce a variation in Va by +0.4% (FHWA 2010), thereby avoiding ignore data sensitive to the model performance. In addition, the study of (Meyer 2012) has shown a comparative approach between RMSE and the standard deviation (SD) which was necessarily adopted in this research to evaluate the model performance with the reference of the standard deviation of acceptance of AASTHO T209 and ASTM D2041 for Rice test.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(yi - \hat{y})^2} \quad (2.2)$$

Where,

1) yi = actual value
2) $\hat{y}$ = predicted value of yi
3) N = number of samples

2.3.3 The Mean Absolute Error (MAE)

The mean absolute error (MAE) was calculated due to it is one of the most insightful metrics by which an observation of the absolute difference between the real values and the model's predictions is achieved. A minor value of MAE indicates the model's predictions are acceptable, in contrast a large value of MAE implies that the model probably has problems in certain parts to adequately produce logical predictions.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|yi - \hat{y}| \quad (2.3)$$

Where,

1) yi = actual value
2) $\hat{y}$ = predicted value of yi

3) N = number of samples

2.4 Validation testing (phase 3)

In the phase 3, the validation dataset of 166 pairs which was built from HMA design of other projects completed in Quebec and Mexico were employed to check the two developed DNN models, thus avoiding models overfitted. In addition, only input data were considered into the regression optimum models (DNN). In effect, these new data (unseen by the models) were used into the models previously trained to obtain the Gmb and Gmm predicted. After, a quality evaluation of the model's performance was implemented by comparing the R^2 obtained using this validation dataset with the same performance metric (R^2) of the optimum model trained. Thus, a model was considered overfitted if R^2 value obtained after validation, is largely lower compared to the R^2 of the optimum model in the selection process. Also, the grade of overfitting was evaluated as how the R^2 validation differs from the R^2 obtained in the development process. However, a model is considered being able to generalize a phenomenon when the coefficient of correlation (R^2) shows a strong correlation over 0.75 after validation. This validation approach has been applied by researchers. For instance, (Feipeng Xiao 2009) used stiffness values, from other projects completed at Clemson University by the Asphalt Rubber Technology Services, to validate the statistical regression and the developed ANN models. Also, the work of (Sebaaly et al., 2018) compared the predictions values of the model developed with laboratory measured values.

2.5 Sensitivity Analysis (correlation analysis)

A sensitivity analysis was performed by removing the input parameters only into the model in which generalization was validated using unseen data, thus avoiding drawing conclusion under a model overfitted. These input variables were removed from the optimum model validated as shown in Table 2.3; three of the four gradation factors were removed by leaving only one gradation factor as shown in Figure 2.11 . Also, the materials (aggregate and binder) inputs were removed one by one. Thus, an evaluation of seven (7) models was achieved with different

inputs variable combinations. All the seven (7) models were trained in the same conditions as the model validated regarding data-split, hyperparameters, neurons and hidden layers. A first evaluation of all seven models was made by using the performance metrics. However, a second evaluation was included by testing the seven (7) models with the validation data set (unknown data), following the process mentioned in the section 2.4 of this chapter. This second test (validation) using unknown data was performed with the idea to increase the experiment's external validity, thus avoiding making a sensitivity analysis on an overfitted model and enhancing the results by which they can be generalized and applied to the broader world. The performance metrics R^2, RMSE and MAE were used to evaluate the degree of correlation strength between each input and the output (Gmm).

Table 2. 3 Shows the changes regarding the inputs variables of the neural network architecture Model 8 selected to simulate Gmm

Number of inputs	Inputs removed	Outputs
8 (original model 8)	Any	Gmm-Gse
5 (Only TG)	GI, G/S, Fp	
5 (Only GI)	TG, G/S, Fp	
5 (Only G/S)	GI, TG, Fp	
5 (Only Fp)	GI, TG, G/S	
7	Gsb	
7	Visc 275	
7	Gb	

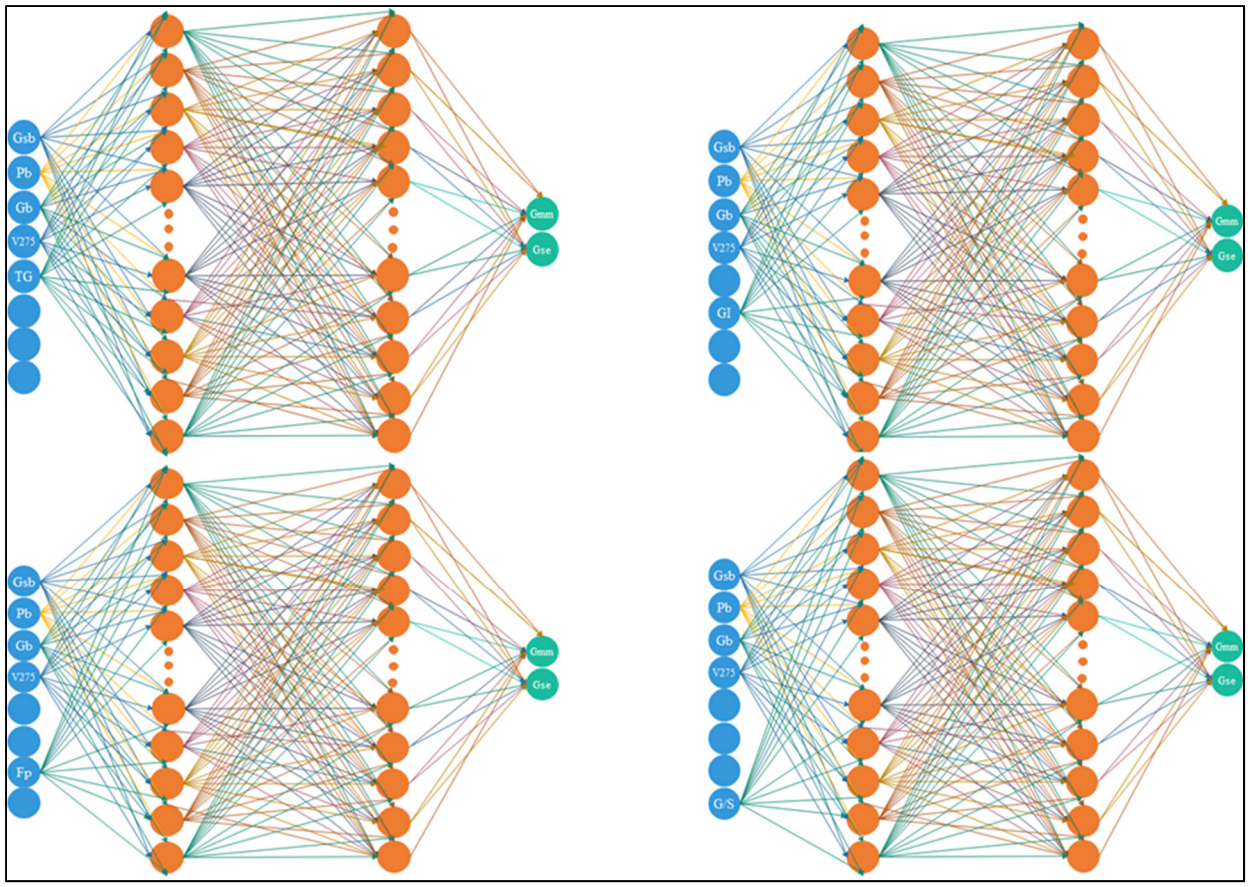


Figure 2. 11 Illustration of the neural network architecture in which gradation factors were changed to evaluate the impact on the Gmm.

CHAPTER 3

EXPERIMENTAL RESULTS

This chapter first presents the three (3) datasets. Two (2) were used to train the 2 DNNs and one (1) small dataset, referred to as the unknown dataset, was used to validate the models previously trained. Second, the models selected are presented based on the correlation coefficient R2 and learning curves to assess generalization. Third, the validation of the selected models to avoid wrong conclusion based on an overfitting model by clarifying whether the model selected could generalize or not the phenomenon studied is presented. Fourth, the correlation analysis for the model validated to evaluate the influence of aggregates and binder properties on the outputs (Gmb or Gmm) is shown. In addition, the prediction performance of the different models was presented supported on error measurement metrics mean average error (MAE) and root mean square error (RMSE).

3.1 Datasets

The datasets were built collecting data from Infopave website according to the procedure described in section 2.1 of chapter 2. The first, the dataset used in the model SGC and predict Gmb values contained 26,620 pairs. This data was collected based in HMA Superpave mix design. The second, dataset used in the Rice test model to predict Gmm values included data of Marshall and Superpave HMA design.

3.1.1 Data pairs for SGC Model

Following the steps described in chapter 2, the raw data collected was processed to build the SGC dataset.

From 87,474 pairs collected, only the 30.43% (equivalent to 26,620 pairs) as shown in Figure 3.2 were retained into the SGC dataset due to:

a) 34.97 % (equivalent to 30,587 pairs) were duplicated. It indicates that the pair has the same values as shown in figure 3.1;

1.026	2.519	2.442	1.1	4.3	14.8	4.8	67.7	2.747	2.664	100	100	100	100	91	62	32	14	9	8	624	5190
1.026	2.519	2.442	1.1	4.3	14.8	4.8	67.7	2.747	2.664	100	100	100	100	91	62	32	14	9	8	642	5176
1.026	2.519	2.442	1.1	4.3	14.9	4.9	67.3	2.747	2.664	100	100	100	100	91	62	32	14	9	8	306	1134
1.026	2.519	2.442	1.1	4.3	14.9	4.9	67.3	2.747	2.664	100	100	100	100	91	62	32	14	9	8	305	1132
1.026	2.519	2.442	1.1	4.3	14.9	4.9	67.3	2.747	2.664	100	100	100	100	91	62	32	14	9	8	644	5345
1.026	2.519	2.442	1.1	4.3	14.9	4.9	67.3	2.747	2.664	100	100	100	100	91	62	32	14	9	8	624	5190
1.026	2.519	2.442	1.1	4.3	14.9	4.9	67.3	2.747	2.664	100	100	100	100	91	62	32	14	9	8	642	5176
1.026	2.519	2.446	1.1	4.3	14.6	4.6	68.8	2.747	2.664	100	100	100	100	91	62	32	14	9	8	306	1134
1.026	2.519	2.446	1.1	4.3	14.6	4.6	68.8	2.747	2.664	100	100	100	100	91	62	32	14	9	8	305	1132
1.026	2.519	2.446	1.1	4.3	14.6	4.6	68.8	2.747	2.664	100	100	100	100	91	62	32	14	9	8	644	5345

Figure 3. 1 Illustration of duplicate pairs

b) 34.60 % were incoherent values regarding:

- ✓ HMA volumetric properties (Gmb greater than Gmm); 29.80% (equal to 26,071 Pairs)
- ✓ Aggregates volumetric properties (Gsb greater than Gse, Gsb greater than Gsa or Gse grater than Gsa); 4.64% (equal to 4057 Pairs);
- ✓ Gmb and Gmm and the states of voids (VMA, Va and VFA). Approximately 0.16% (139 pairs)

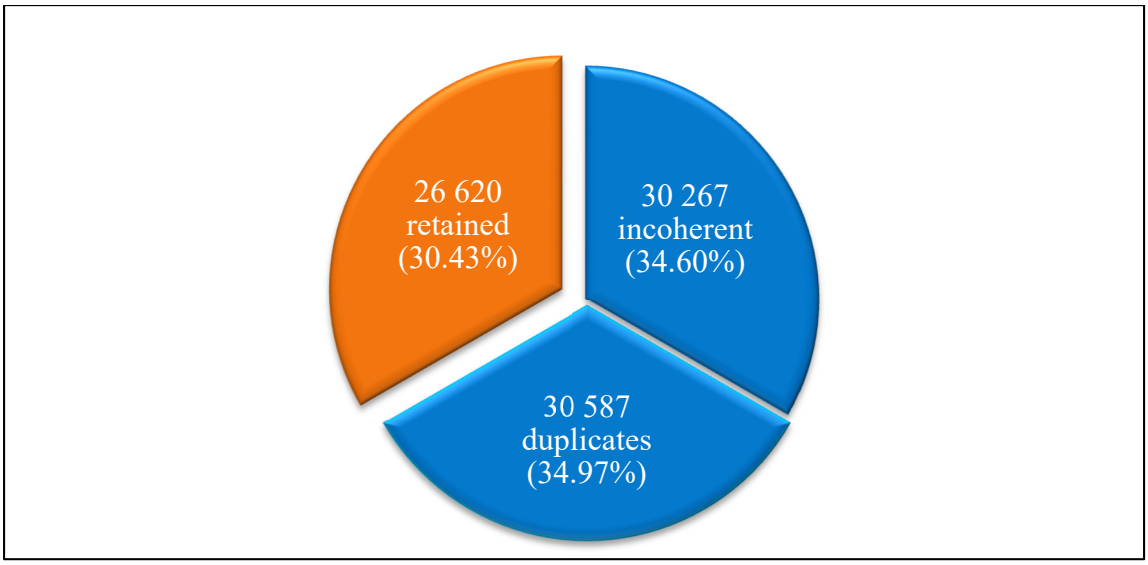


Figure 3. 2 Circular chart that shows the amount of data retained to build the dataset for SGC model after cleaning processing

Table 3.1 shows the list of input and output parameters used in the ANN model to predict Gmb, along with the range of values used in training.

Table 3. 1 Parameters used in the ANN training process to predict Gmb and range of values

Gmb model	Parameter	Min.	Max.	Avg.	SD
Inputs	N_G	7	149	79.098	31.271
	Gsb	2.488	2.887	2.654	0.053
	Gb	1.006	1.044	1.025	0.008
	Pb	3.6	6.8	4.99	0.524
	Visc 275 F	229	1475	551.744	204.693
	TG	285.9	702	497.871	87.985
	GI	8.6	39.4	16.57	6.397
	Fp	0.042	0.304	0.112	0.046
	G/S	0.48	3.90	1.59	0.854
Outputs	Gmb	2.229	2.644	2.401	0.068

3.1.2 Data Pairs for Rice Test Model

The raw data collected to Predict Gmm test was processed as described in chapter 2. This raw data showed the following characteristics:

1) From 376,128 pairs collected, 316,594 (84.17%) as shown in Figure 3.3 were retained to be part of the Rice dataset because
   a) 49,786 pairs (13.24%) were duplicated or also missing values, thus becoming difficult the calculation of the missed values to complete the pair
   b) Incoherent values regarding:
      - ✓ HMA volumetric properties (Gmb greater than Gmm); 7,627 pairs (2.03%)
      - ✓ aggregates volumetrics properties (Gsb greater than Gse, Gsb greater than Gsa or Gse grater than Gsa); 2,121 pairs (0.56%)

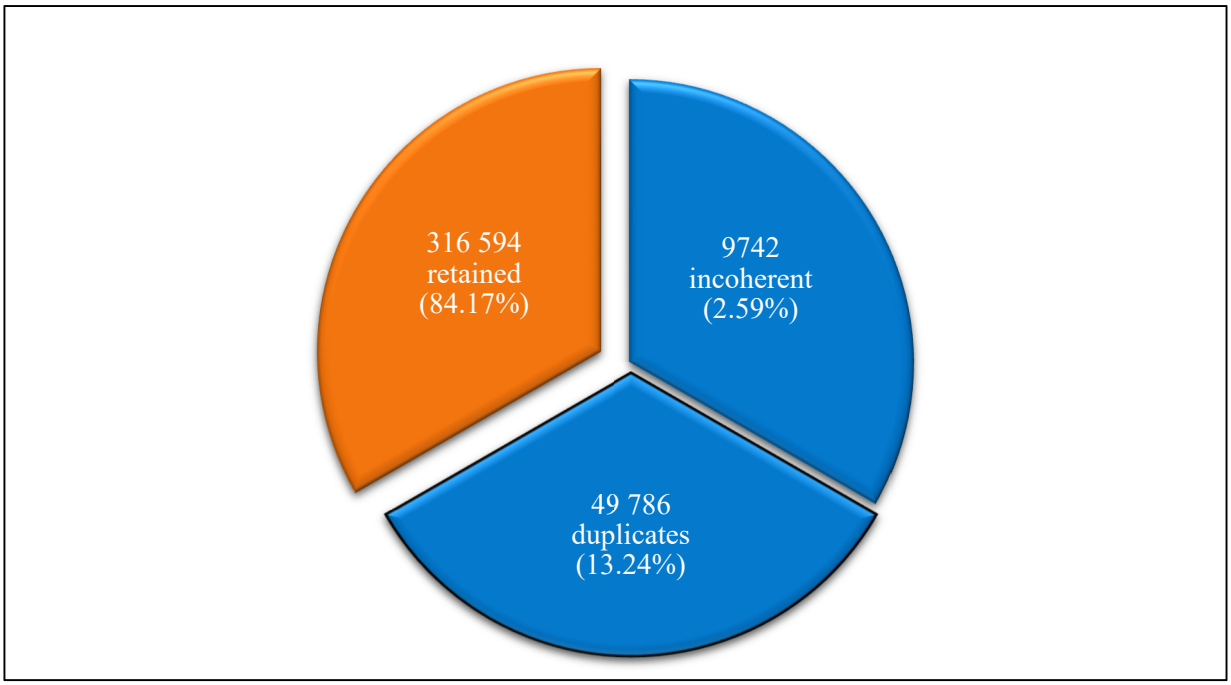


Figure 3. 3 Circular chart that shows the amount of data retained to build the dataset for *Rice test* simulation after cleaning processing

The normal distribution of input data regarding Pb, Gb, Gsb and Viscosity at 275 F is shown in Figure 3.4.

Additionally, the Figure 3.5. has the normal distribution of the four quantification gradation factors (FP,G/s, TG and GI) as inputs.

The table 3.2 provides information concerning the variability of data collected in which the Rice test models were training.

Table 3. 2 Parameters used in the ANN training process to predict Gmm and range of values

Gmm model	Parameter	Min.	Max.	Avg.	SD
	Gsb	2.485	3.000	2.664	0.1095
	Gb	1.002	1.107	1.026	0.012
	Pb	2.4	7.5	4.85	0.708

Inputs	Visc 275 F	100	19406	874	590.570
	TG	144.2	872	442.844	106.422
	GI	3.2	40.8	15.475	7.389
	Fp	0.025	1.65	0.138	0.075
	G/S	0.030	33.57	1.801	1.923
Outputs	Gmm	2.334	2.781	2.527	0.089
	Gse	2.558	3.145	2.732	0.124

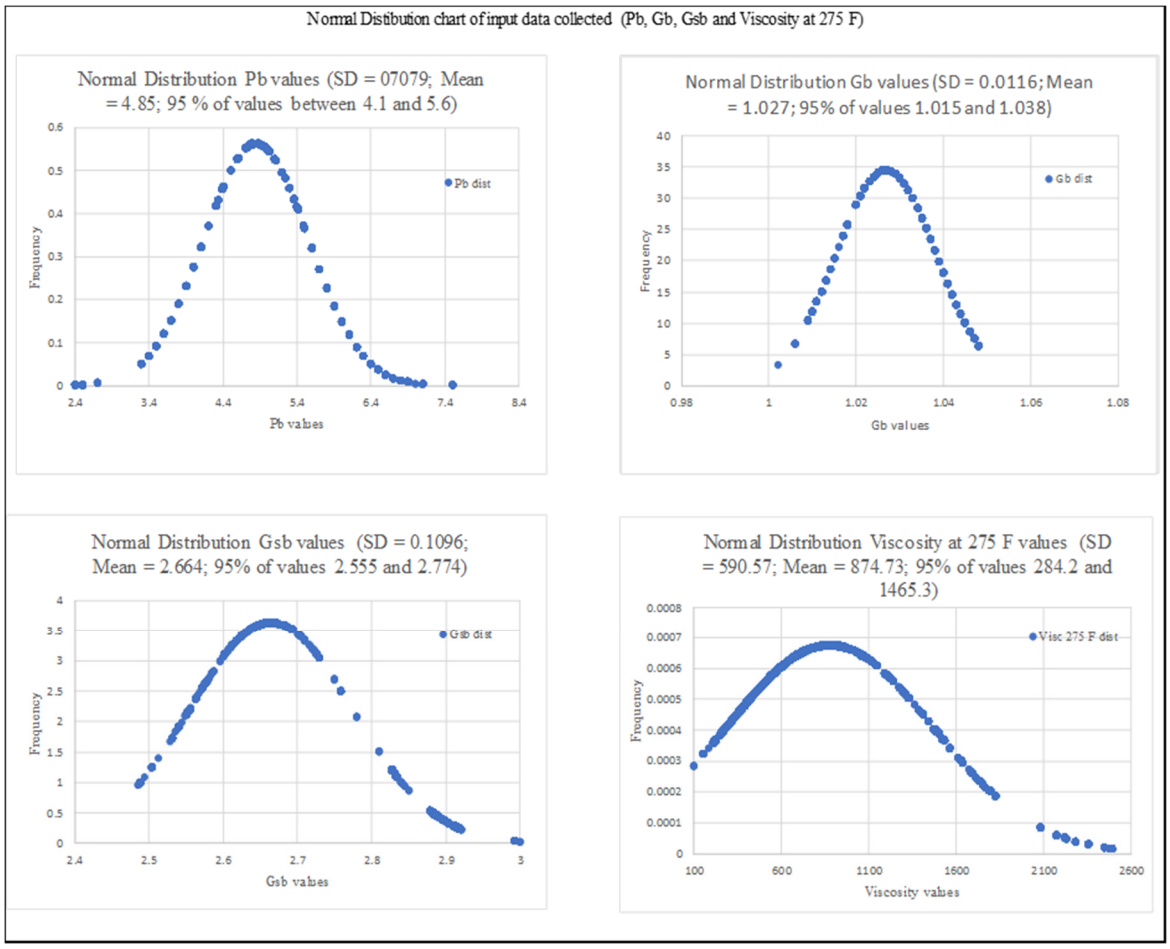


Figure 3. 4 Normal distribution charts for each input data collected regarding Pb, Gb, Gsb and Viscosity at 275 F

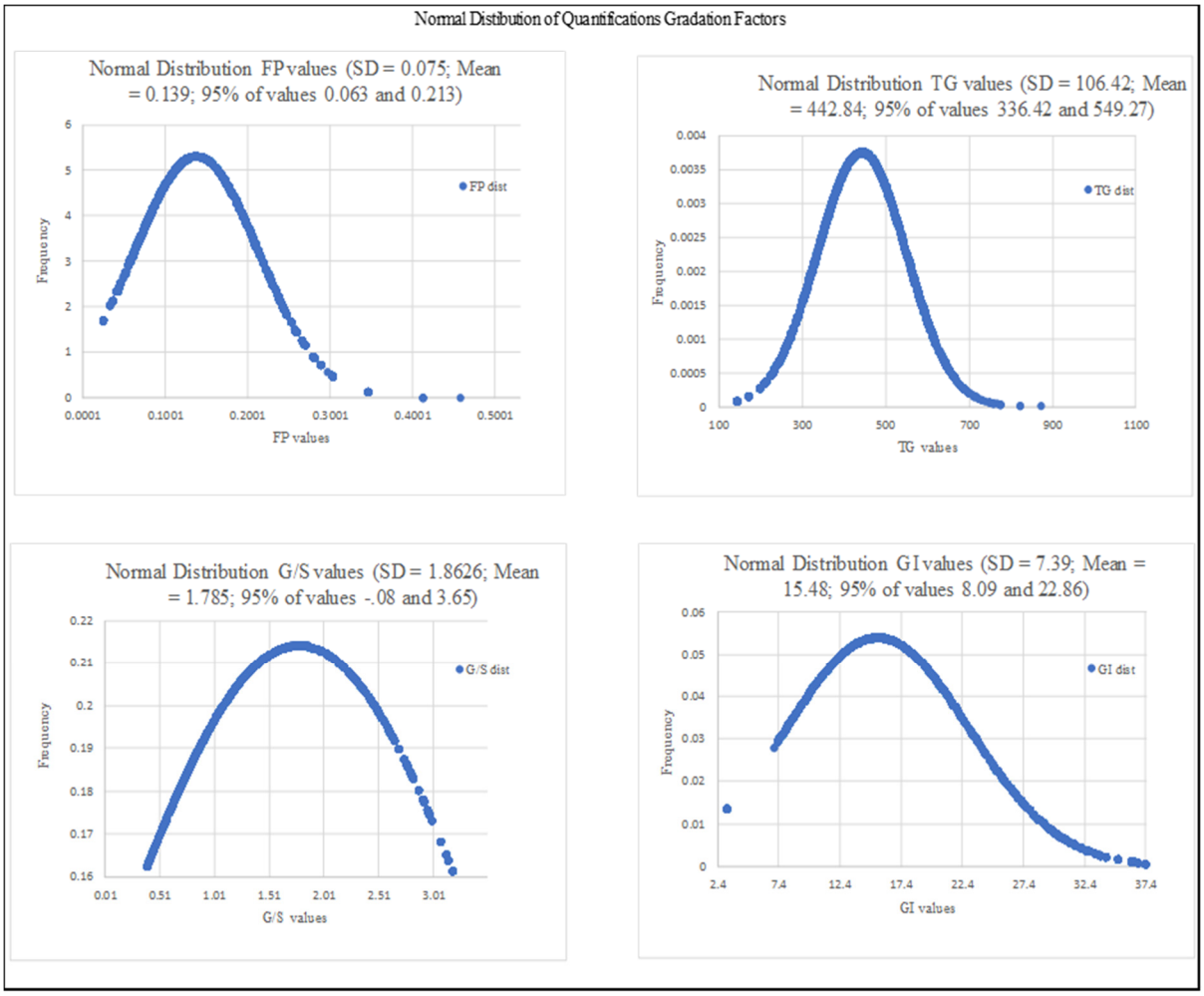


Figure 3. 5 Normal distribution charts for each input data collected regarding the quantification factors (FP, TG, G/S and GI)

A clear normal distribution curve is seen in Figure 3.5 for the four gradation factors, except G/S. These curves show different shapes despite all four factors were calculated from the same gradation

3.1.3 The Validation Dataset

This dataset presented no problem regarding duplication or any of the incoherence found because it was manually built directly from laboratory reports for HMA design and certifications from material suppliers regarding material properties. This data was collected

from HMA designed in Mexico and Quebec, thus unknowing by the models trained with the secondary data from Infopave. Table 3.3 shows the list of parameters used to validate Gmb and Gmm ANN-models, along with the range of values used in this process.

Table 3. 3 the range of values used in the validation process

	Parameter	Min.	Max.	Avg.	SD
Inputs	N_G	6	205	83.15	0.003
	Gsb	2.678	2.833	2.700	0.037
	Gb	1.017	1.030	1.021	0.004
	Pb	4.0	5.69	5.265	0.358
	Visc 275 F	528.38	1378	1122.81	180.717
	TG	401.671	484.564	443.117	41.447
	GI	12.45	26.79	17.04	3.133
	Fp	0.037	0.265	0.145	0.059
	G/S	0.050	1.490	0.973	0.263
Outputs	Gmb	2.016	2.662	2.331	0.120
	Gmm	2.464	2.775	2.514	0.061
	Gse	2.708	2.986	2.737	0.065

3.2 DNN Models Selection for Gmm and Gmb Prediction

In accordance with the procedure specified in the chapter 2 in which empirically the DNN architecture were tuned by changing the hyper-parameters (learning rate, dropouts, number of layers and nodes) to test the hypothesis of Gmb and Gmm test can be predicted using a DNN as shown in Tables 3.4 and 3.5. Table 3.4 shows the DNNs developed to simulate the SGC test by predicting Gmb values. These DNNs for Gmb prediction model contained nine (9) independent variables, the gyrations numbers, aggregate properties (Gsb, and four Gradations factors), binder properties (viscosity, Gb and Pb) as inputs, and bulk specific gravity of the HMA (Gmb) as output.

Table 3. 4 Correlation coefficient R2 results from DNNs developed to SGC test model by predicting Gmb. These models were developed using a learning rate of 0.0001 and batch of 10

DNN for Model Selection (Hyperparameter changes)						
Training (Model)	Outputs	% training	Layers	Nodes	Dropout	R2
1	Gmb	75	4	150	0.15	0.9554
2	Gmb	75	4	200	0.15	0.9608
3	Gmb	75	5	350	0.15	0.9533
4	Gmb	70	5	450	0.15	0.9580
5	Gmb	70	6	450	0.15	0.9586
6	**Gmb**	**70**	**5**	**350**	**0.10**	**0.9617**

The DNNs developed to predict Gmm under Rice test are composed by eight 8 of the nine 9 independent variables mentioned before because the gyrations number was not considered in this DNN architecture. During the selection process of the Rice test simulation, 2 types of DNNs were tested. As can be seen in Table 3.5, models 2, 3, 4, 5, 6 and 7 were trained with only Gmm as outputs and models 1 and 8 were trained two of them (Gmm and Gse).

Table 3. 5 Correlation coefficient R2 results from DNNs developed to Rice test model by predicting Gmm. These models were developed using a learning rate of 0.0001 and batch of 10

DNN for Model Selection (Hyperparameter changes)						
Training (Models)	Outputs	% training	Layers	Nodes	Dropout	R2
1	Gmm-Gse	70	3	150	0.15	0.9837
2	Gmm	70	3	150	0.15	0.9746
3	Gmm	70	2	150	0.15	0.9875
4	Gmm	60	2	150	0.15	0.9873
5	Gmm	70	2	250	0.15	0.9886
6	**Gmm**	**70**	**2**	**150**	**0.15**	**0.9869**
7	Gmm	60	2	150	0.2	0.9829
8	**Gmm_Gse**	**70**	**2**	**150**	**0.15**	**0.9887**

The models with the highest correlation coefficient R^2 for Gmb and the Gmm and also based on learning curve which (indicates behaviours of overfitting or underfitting, as well as the error measure (RMSE), were selected.

Moreover, the Tables 3.4 and 3.5 show the results of different training performed in which hyper-parameters were changed following a basic search grid scheme (this scheme allows

covering different combinations of parameters in an orderly way). Table 3.4 corresponds to the R^2 result for different neural architectures regarding Gmb prediction model. According to the Table 3.4 when the number of layers and neurons were increased into the DNN, from 4 to 5 and nodes from 150 to 350, the R^2 increased from 0.9554 to 0.9617. However, when hidden layers were increased by 6 (model 5), R^2 reduced to 0.9586. As can be seen in Table 3.4, the model 6 trained with split data of 70 % for training and 30 % for testing obtained the highest R^2 equal to 0.9617. Additionally, any overfitting problem was observed as can be seen in the Figures 3.5, 3.6 and 3.7 regarding learning curves in which the loss curve of all models was almost always lower on the training than the test dataset. The curve (orange) of test loss decreases steadily and then stabilizes after epoch 30 for the model 1 (Figure 3.6), epoch 20 for models 4 and 6 as shown in Figures 3.7 and 3.8 with a little gap regarding the training loss curve (blue) which suggests only a very small overfit. The representation learning curve of models 2, 3 and 5 can be seen in annexe I.

DNN Architecture of the Model 1 for SGC simulation						
Training (Model 1)	Outputs (Gmb)	% traininig (75)	Layers (4)	Nodes (150)	Dropout (0.15)	R2 (0.9554)

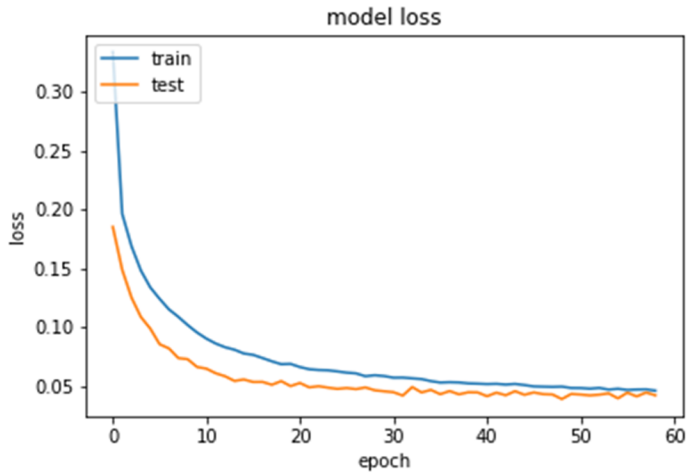


Figure 3. 6 Learning curves of model 1 to predict Gmb values. The curve (orange) of test loss decreases steadily and then stabilizes after epoch 30. The training loss curve (blue)

DNN Architecture of the Model 4 for SGC simulation						
Training (Model 4)	Outputs (Gmb)	% traininig (70)	Layers (5)	Nodes (450)	Dropout (0.15)	R2 (0.9580)

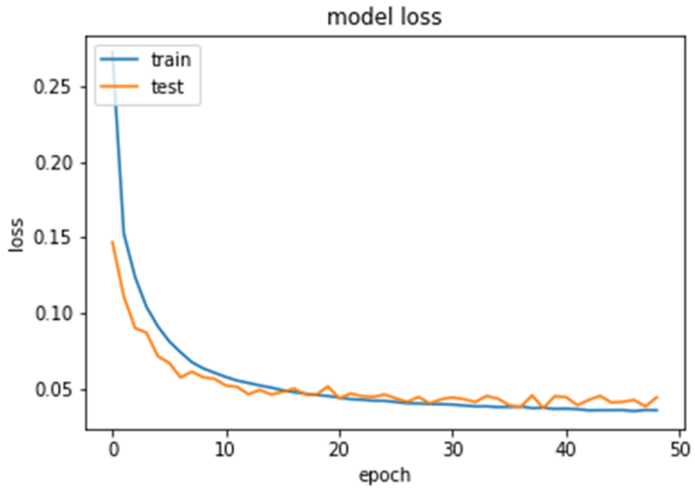


Figure 3. 7 Learning curves of model 4 to predict Gmb values. The curve (orange) of test loss decreases steadily and then stabilizes after epoch 20. The training loss curve (blue)

DNN Architecture of the Model 6 for SGC simulation						
Training (Model 6)	Outputs (Gmb)	% traininig (70)	Layers (5)	Nodes (350)	Dropout (0.10)	R2 (0.9617)

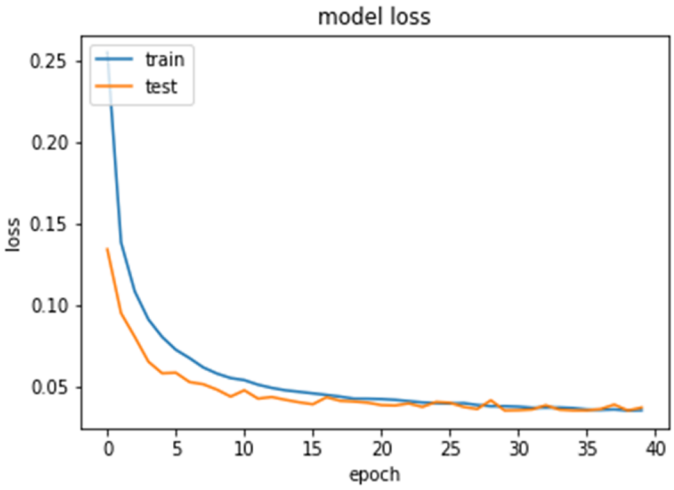


Figure 3. 8 Learning curves of model 6 to predict Gmb values. The curve (orange) of test loss decreases steadily and then stabilizes after epoch 20. The training loss curve (blue)

The Table 3.5 shows the performance of Gmm prediction models based on R^2. The highest R^2 is indicated in Table 3.5 (model 8). In addition, despite DNNs architectures with more than 2 to hidden layers obtained R^2 values equal to 0.9837 and 0.9746, overfitting problems were observed in which the curve of training loss (blue) remained decreasing, however, the test curve (orange) increased in a point without stability over the training curve (blue). This behaviour was observed for models 1 and 2 as can be seen in Figures 3.8 and 3.9. For instance, the test curve increased after the epoch 5 for the model 1 (Figure 3.9) and after the first epoch for model 2 (Figure 3.10), over the training curve without stability.

DNN Architecture of the Model 1 for Rice test simulation						
Training (Model 1)	Outputs (Gmm-Gse)	% training (70)	Layers (3)	Nodes (150)	Dropout (0.15)	R2 (0.9837)

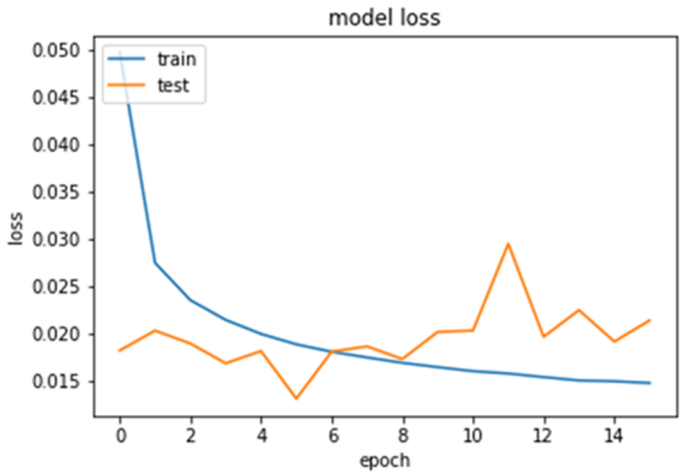


Figure 3. 9 Learning curves of model 1 to predict Gmm values. The curve (orange) of test loss decreases steadily and then increases after epoch 5. The training loss curve is represented in blue

DNN Architecture of the Model 2 for Rice test simulation						
Training (Model 2)	Outputs (Gmm)	% training (70)	Layers (3)	Nodes (150)	Dropout (0.15)	R2 (0.9746)

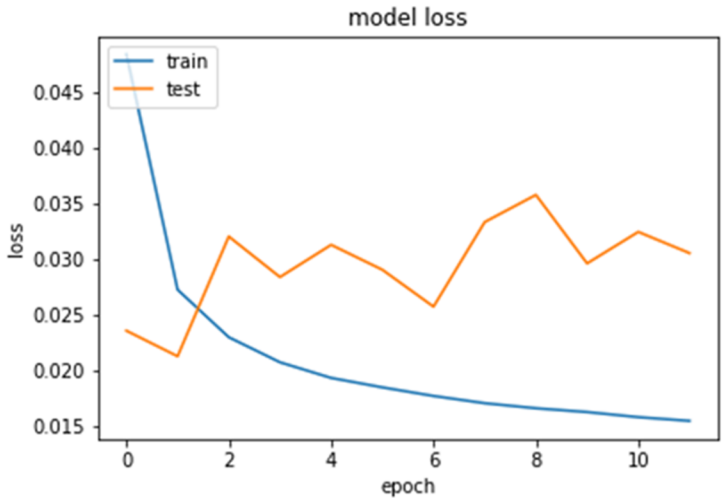


Figure 3. 10 Learning curves of model 2 to predict Gmm values. The curve (orange) of test loss decreases steadily and then increases after the first epoch. The training loss curve (blue)

In fact, for models with 2 hidden layers a good fit learning curves was achieved as predicted by the models 3, 4, 5, 6, 7, and 8 in which the curve (orange) of test loss decreases steadily and then stabilizes after epoch 20. An illustration of this behaviour is shown in Figures 3.11 and 3.12, and 3.13 for models 3, 5 and 7 respectively. The learning curves of models 4 and 6 can be seen in annexe II.

DNN Architecture of the Model 3 for Rice test simulation						
Training (Model 3)	Outputs (Gmm)	% training (70)	Layers (2)	Nodes (150)	Dropout (0.15)	R2 (0.9875)

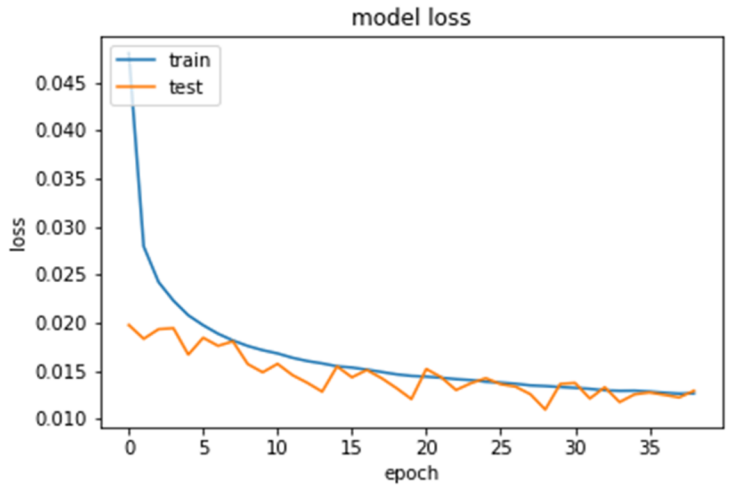


Figure 3. 11 Learning curves of model 3 to predict Gmm values. The curve (orange) of test loss decreases steadily and then stabilizes after epoch 20. The training loss curve (blue)

DNN Architecture of the Model 5 for Rice test simulation						
Training (Model 5)	Outputs (Gmm)	% training (70)	Layers (2)	Nodes (250)	Dropout (0.15)	R2 (0.9886)

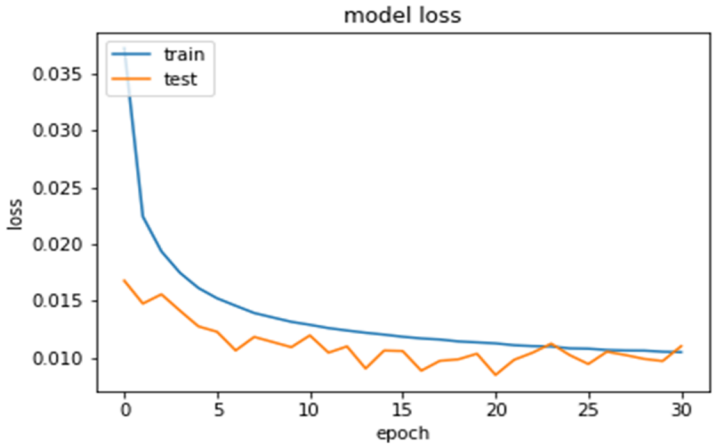


Figure 3. 12 Learning curves of model 5 to predict Gmm values. The curve (orange) of test loss decreases steadily and then stabilizes after epoch 20. The training loss curve (blue)

DNN Architecture of the Model 7 for Rice test simulation						
Training (Model 7)	Outputs (Gmm)	% training (60)	Layers (2)	Nodes (150)	Dropout (0.20)	R2 (0.9829)

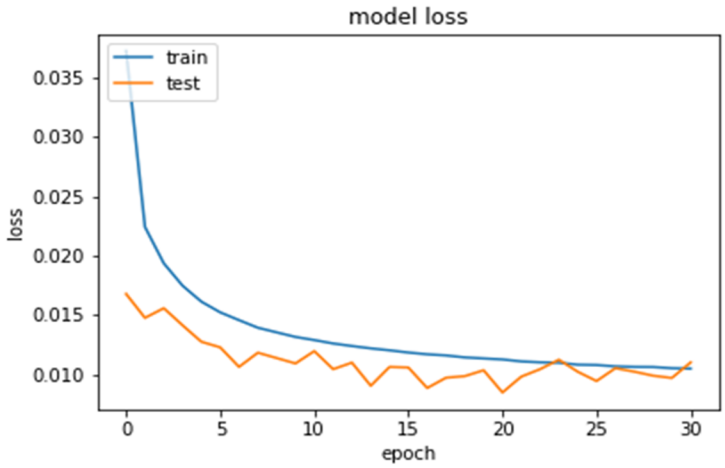


Figure 3. 13 Learning curves of model 7 to predict Gmm values. The curve (orange) of test loss decreases steadily and then stabilizes after epoch 20. The training loss curve (blue)

Additionally, an optimization of the model was observed from models 7 to 8 when 2 outputs (Gmm and Gse) were added to the DNN architecture instead of training models with Gmm as the only output, thus increasing R^2 from 0.9869 to 0.9887. The learning curves as can be seen in Figure 3.14 showed that the amount of test equivalent to 30 % of the dataset may be too small relative to the training dataset.

DNN Architecture of the Model 8 for Rice test simulation						
Training (Model 8)	Outputs (Gmm-Gse)	% training (70)	Layers (2)	Nodes (150)	Dropout (0.15)	R2 (0.9887)

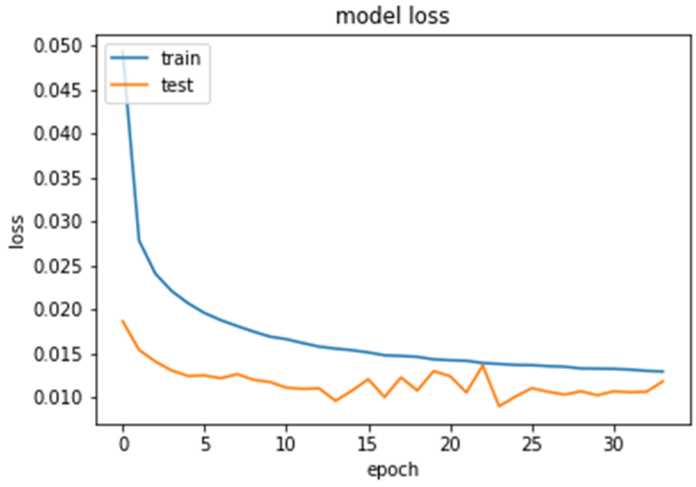


Figure 3. 14 Learning curves of model 8 to predict Gmm values. The curve (orange) of test loss decreases steadily and then stabilizes after epoch 10. The training loss curve (blue)

3.3 Model 6 of the SGC-DNN to Predict Gmb

The DNN based on 9 inputs (gyrations numbers, aggregate properties (Gsb, TG, GI, G/S, Fp), binder properties (viscosity at 275 F, Gb and Pb) and Bulk specific gravity of the HMA (Gmb) as the only output, referred to model 6 allowed validate the hypothesis of SGC can be simulated to predict Gmb values. This model 6 had a mean average error (MAE) of 0.0077 and root mean square error (RMSE) of 0.0127 as shown in the Table 3.6.

Table 3. 6 Performance metric (MAE, RMSE and R^2) of model 6 to simulate SGC by predicting Gmb values

Model 6 (SGC-Gmb)		
MAE	**RMSE**	**R2**
0.0077	0.0127	0.9617

In addition, the normal distribution of real Gmb values and Gmb predicted is illustrated in Figure 3.15 which show that the distribution of Gmb predicted values (red curve) is superposed on the distribution of real Gmb values (blue curve). Both curves with a mean average of 2.401 and almost the same SD of 0.065 for the prediction curve and SD equal to 0.067 for the real values curve. These normal distribution curves were displayed to evaluate differences between real and prediction values under the same dataset

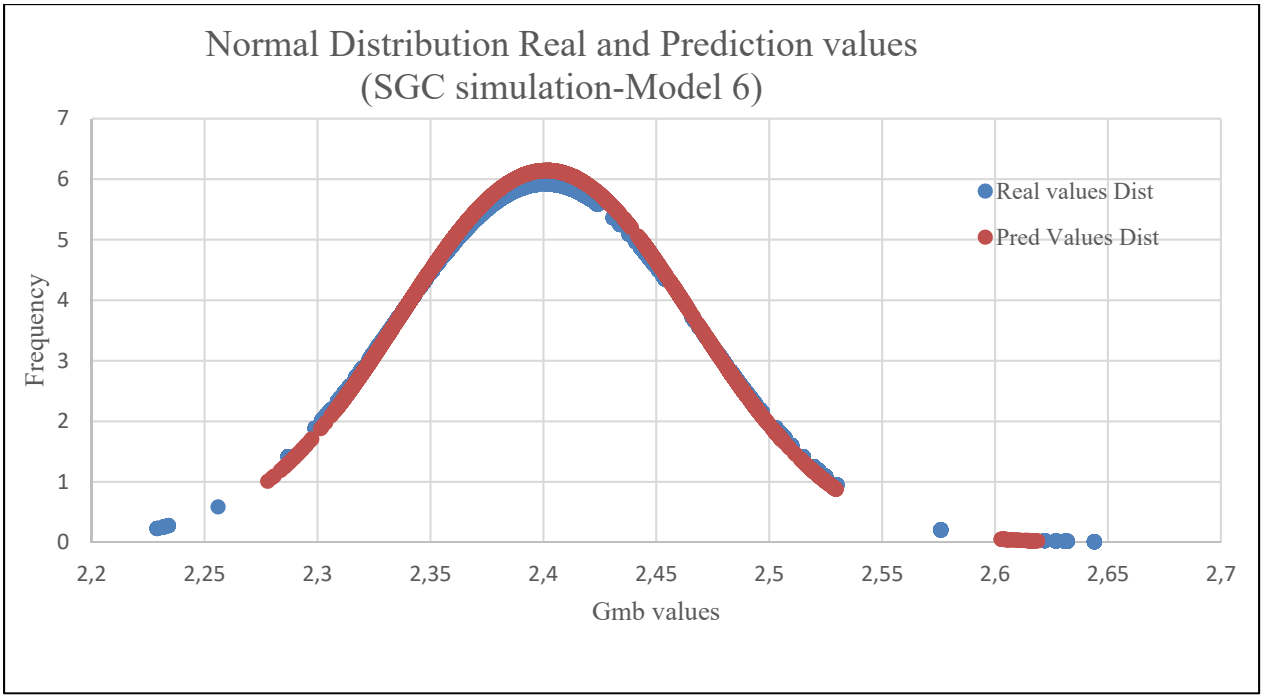


Figure 3. 15 Normal distribution curves of real and prediction values of model 6 for the SGC simulation to predict Gmb values

Figure 3.16 shows a linear correlation of the model 6, between Gmb predicted and Gmb real in which the coefficient R^2 is 0.9617 as shown in table 3.6.

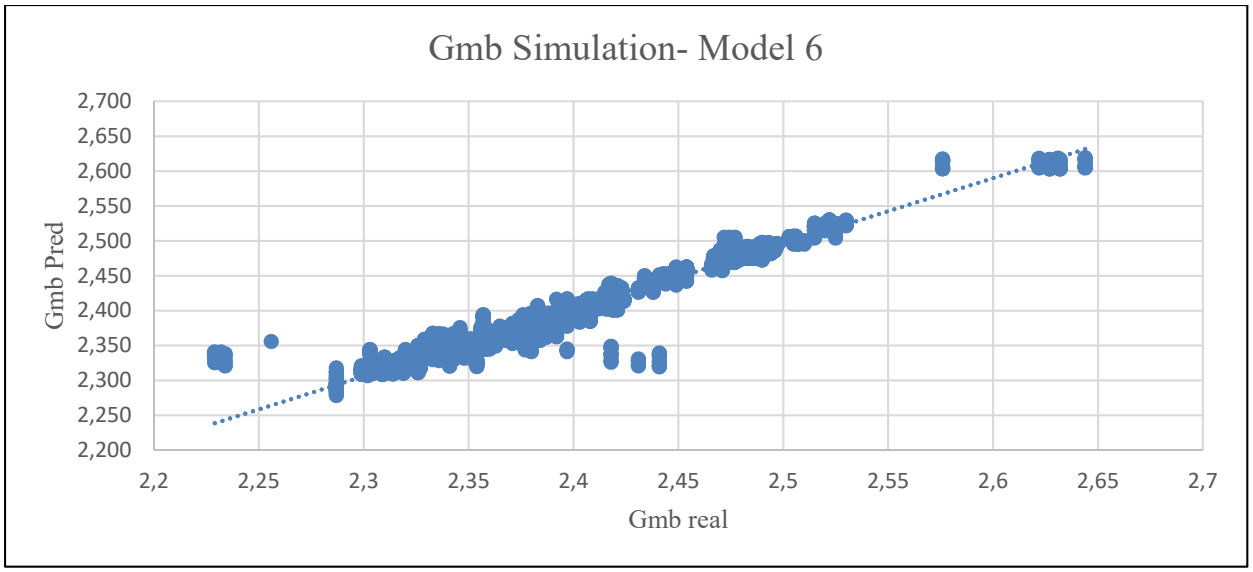


Figure 3. 16 Graphic of linear correlation between real and predicted values of Gmb of the model 6

The Figure 3.17 presents the prediction behaviour during the training from epoch 1 to epoch 29, showing the continuity and precision of the prediction values.

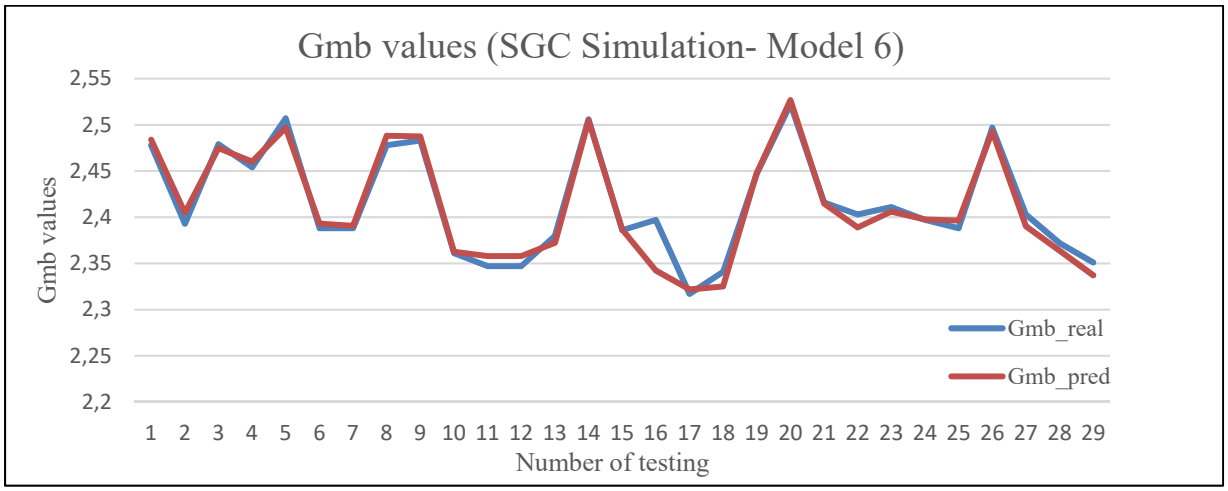


Figure 3. 17 Illustration of correlation between real and predicted values of Gmb of the model 6 during the training process

3.4 Model 8 Rice test to predict Gmm

The DNN architecture of model 8 in which 8-input variables (Gsb, TG, GI, G/S, Fp, kinematic viscosity at 275 F, Gb and Pb) and 2 outputs the maximum specific gravity of the HMA (Gmm) and effective specific gravity of aggregate (Gse) were used, permitted to test the hypothesis of rice test simulation can be achieved to predict Gmm, if enough amount of data pairs is available.

Table 3. 7 Performance metric (MAE, RMSE and R^2) of model 8 to simulate the Rice test by predicting Gmm values

Model 8 (Rice)		
MAE	**RMSE**	**R^2**
0.0065	0.0090	0.9887

Thus, MAE equal to 0.0065 and RMSE equivalent to 0.0090 as shown in the table 3.7. The normal distribution of real Gmm values and Gmm predicted is illustrated in Figure 3.18 in which the distribution of Gmm predicted values (orange curve) is superposed on the distribution of real Gmm values (blue curve). Both curves have a similar mean average of 2.528 for predicted values and 2.527 for real values. Additionally, similar standard deviation values of 0.085 for the prediction curve and 0.089 for the real values curve.

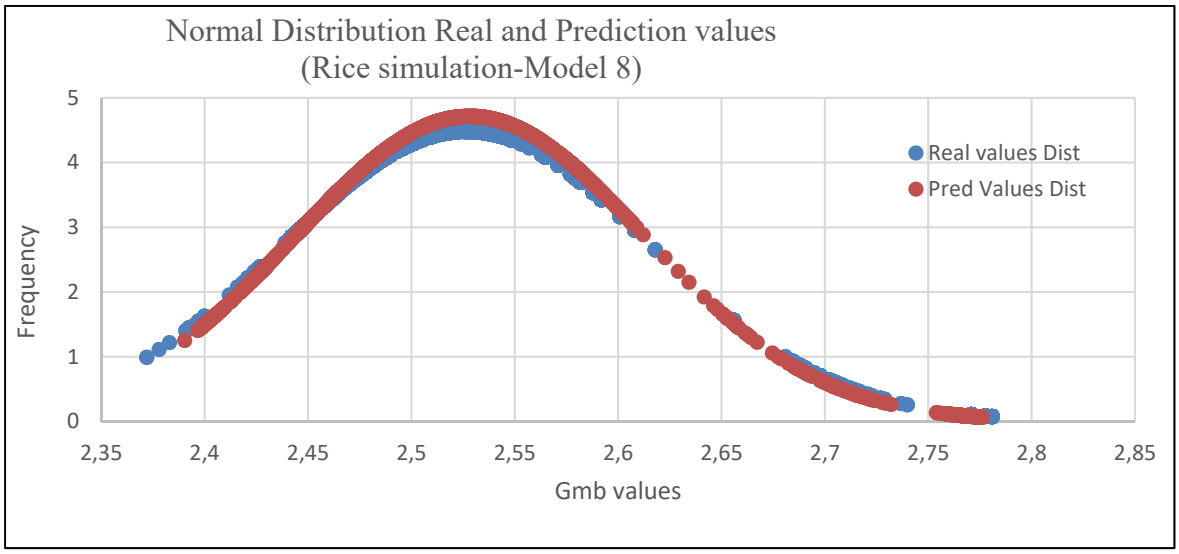


Figure 3. 18 Normal distribution curves of real and prediction values of model 8 for Rice test simulations to predict Gmm values

Figure 3.19 shows a linear correlation of model 8, between Gmm predicted and Gmm real in which the coefficient R^2 is 0.9887 as shown in table 3.7.

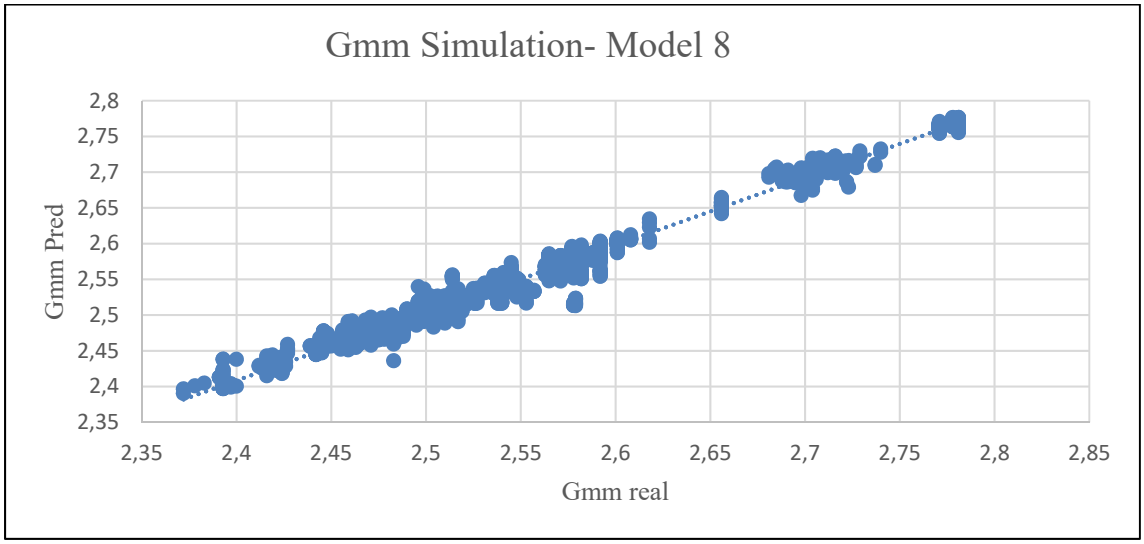


Figure 3. 19 Graphic of linear correlation between real and predicted values of Gmm of the model 8

Figure 3.20 presents the prediction behaviour during the training from epoch 1 to epoch 29, showing the continuity and precision of the Gmm prediction values.

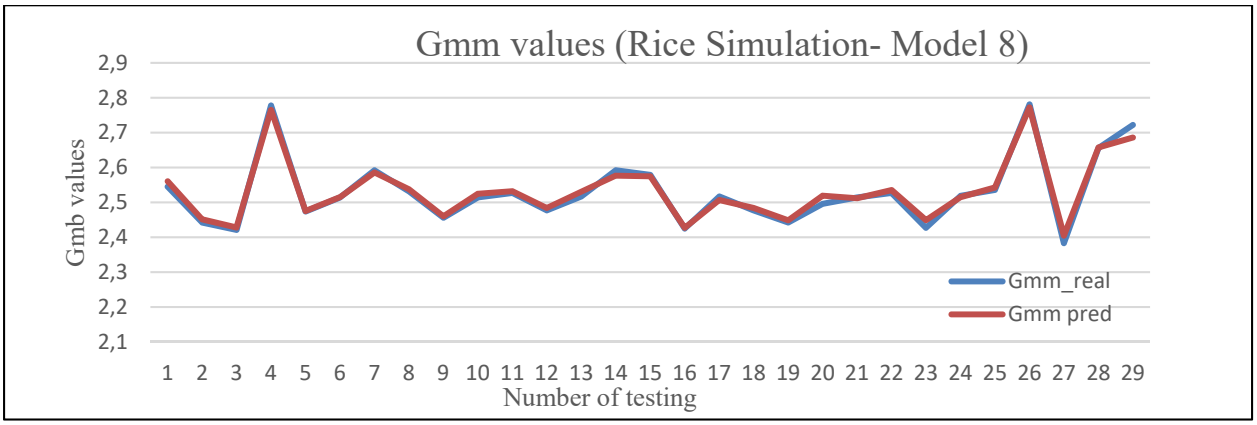


Figure 3. 20 Illustration of correlation between real and predicted values of Gmm of the model 8 during the training process

3.5 Model Validation

Validation was performed using the testing dataset which contains 166 pairs (not seen before by the model). This validation test permitted to further clarify if the model trained could really generalize the phenomenon based on the data available, thereby testing the hypothesis of the SGC and the Rice tests can be simulated to accurately predict Gmb and Gmm values. Moreover, this validation allowed the author to know if the assumption of using a DNN could establish a correlation analysis in which 8 or 9 independent variables can be studied at the same time regarding the phenomenon, thus finding which variables have more influence on the Gmb or Gmm prediction.

3.5.1 Validation of Model 6 (SGC) to Predict Gmb

The external validity concept was applied on Model 6-Gmb. Table 3.8 shows the performance metric (MAE, MRSE and R2). As can be seen in this table 3.9 the correlation coefficient, R2

decreased from 0.9617 to 0.4573. The MAE increased from 0.007 to 0.08447 and MRSE increased from 0.012 to 0.1004.

Table 3. 8 Performance metric (MAE, RMSE and R^2) after validation using unseen data by the model 6 to simulate SGC by predicting Gmb values

Validation - Model 6 (SGC-Gmb)		
MAE	**RMSE**	**R^2**
0.08447	0.10044	0.4573

The Figure 3.21 illustrates the standard distribution of real and predicted values by the model in which a great disparity can be observed between the real (blue curve) and predicted (orange curve) of the Gmb values. The mean average of both curves are similar to 2.329 for real Gmb values and 2.292 for predicted Gmb values. However, the SD of real values is 0.122 and the SD of predicted values is 0.055. As can be seen in Figure 3.21 the model 6 predicts values close to the mean, however, real values are more dispersed.

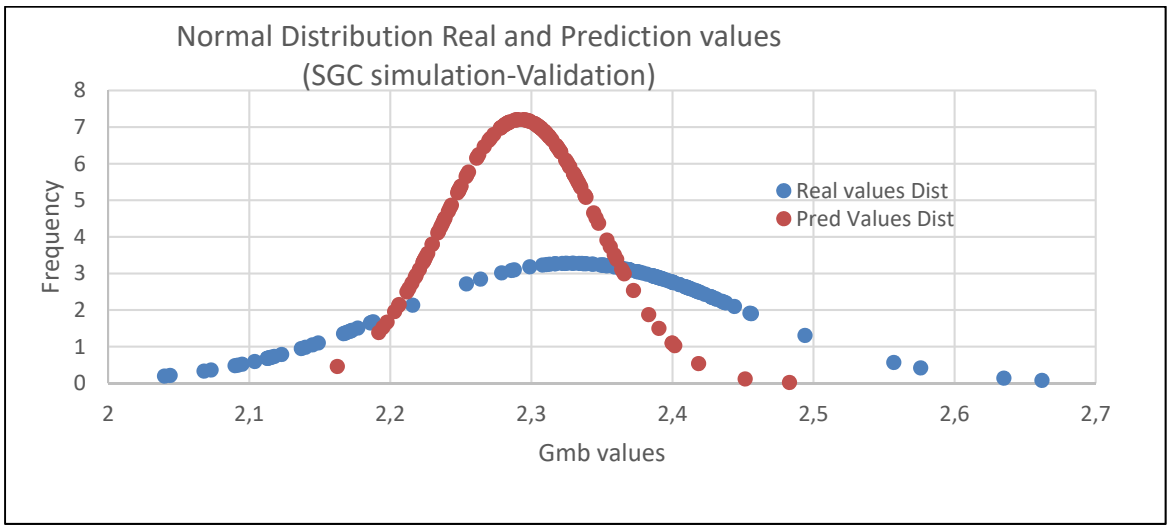


Figure 3. 21 Normal distribution curves of real and prediction values of model 6 using unseen datasets by the model.

Figure 3.21 plots the dispersion of the model 6-Gmb, between Gmm predicted and Gmm real in which the coefficient R2 is 0.4385 as shown in table 3.8.

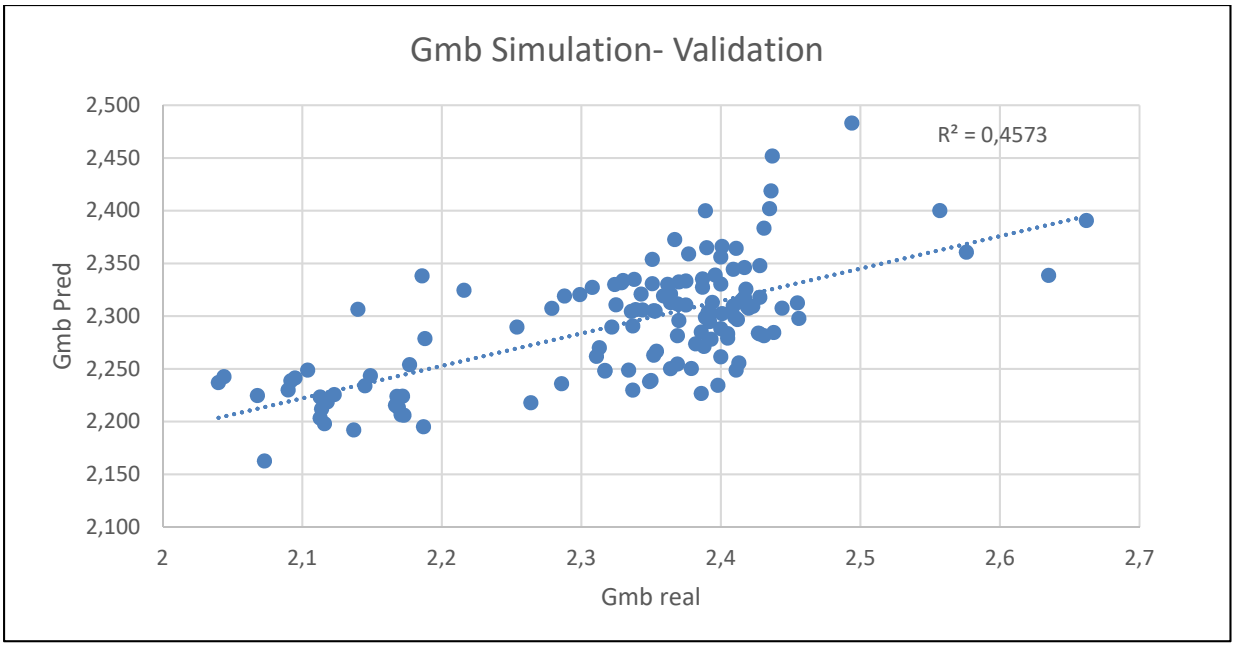


Figure 3. 22 Graphic of the correlation between real and predicted values of Gmb using unseen data by the model 6

Figure 3.23 presents the prediction behaviour validation testing epochs from 1 to 29, showing the discontinuity and inaccuracy of the Gmm prediction values.

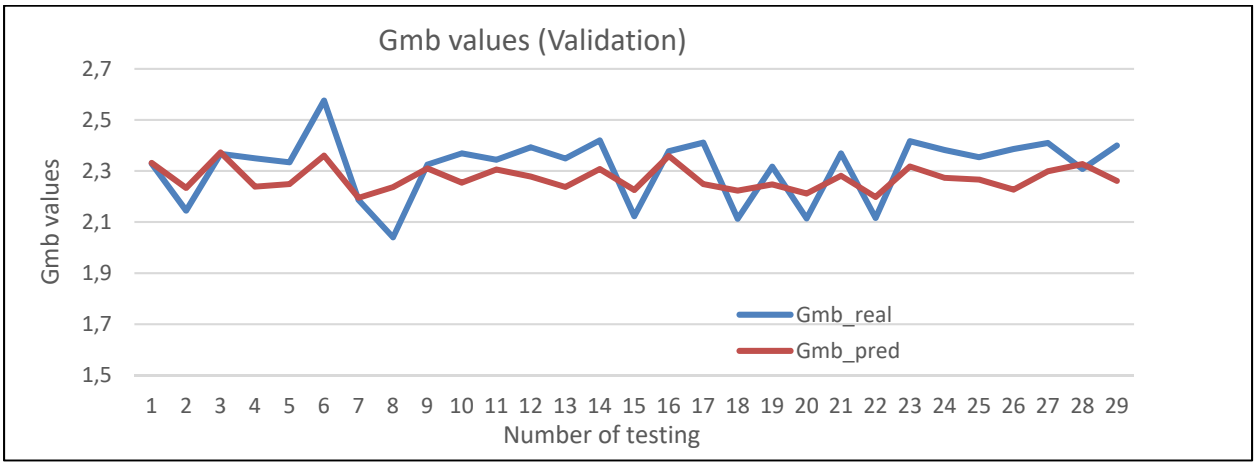


Figure 3. 23 Illustration of correlation between real and predicted values of Gmb of the model 6 during the validation process

3.5.2 Validation of mode 8 Rice Tests to Predict Gmm

Validation was performed on Model 8-Gmm. Table 3.9 shows the performance metric (MAE, MRSE and R^2). As can be seen in tables 3.8 and 3.10, the correlation coefficient R^2 decreased from 0.9887 to 0.9385. The MAE increased from 0.0065 to 0.0080 and MRSE increased from 0.0090 to 0.0130.

Table 3. 9 Performance metric (MAE, RMSE and R2) after validation using unseen data by the model 8 to simulate the Rice test by predicting Gmm values

Validation Model 8 (Rice)		
MAE	**RMSE**	**R^2**
0.0080	0.0130	0.9385

Figure 3.24 illustrates the standard distribution of real and predicted values by the model in which similar curves can be observed between the real (blue curve) and predicted (orange curve) of the Gmm values. Additionally, the mean average (M) and of both curves are similar with 2.514 for real Gmm values and 2.511 for predicted Gmm values. Equally, the SD of real values is 0.0523 and the SD of predicted values is 0.0613. As can be seen in Figure 3.24 the model 8 predicts values close to the mean, as well as the Gmm real values.

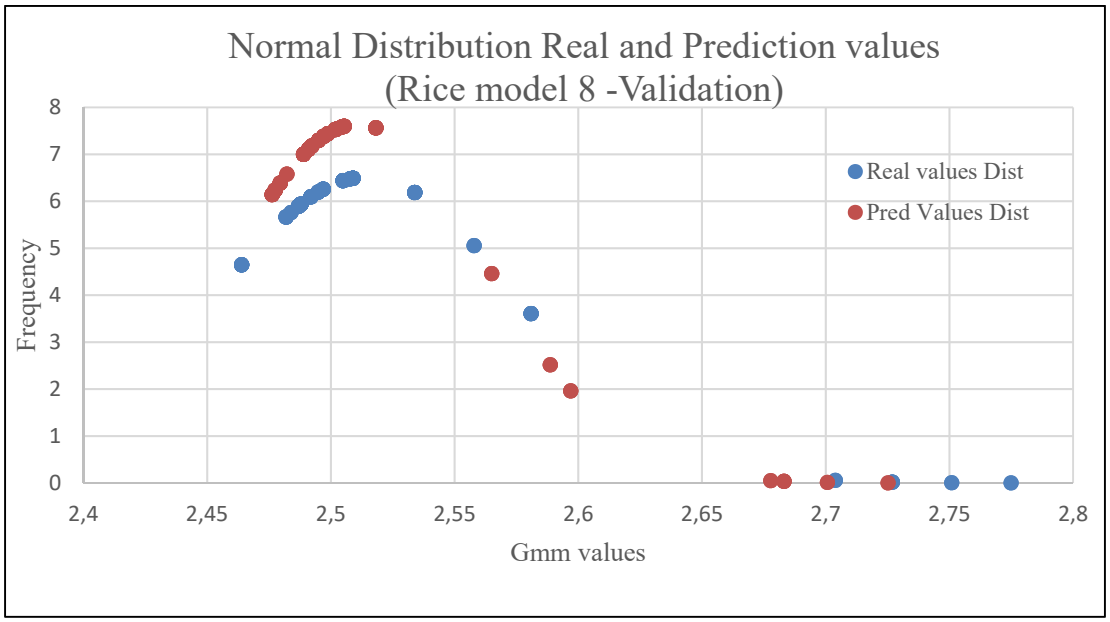


Figure 3. 24 Normal distribution curves of real and prediction values of model 8 using unseen datasets by the model.

Figure 3.25 plots a linear correlation of the model 8, which was validated by 166 new pairs samples, between Gmm predicted and Gmm real in which the coefficient R2 is 0.9385 as shown in table 3.9.

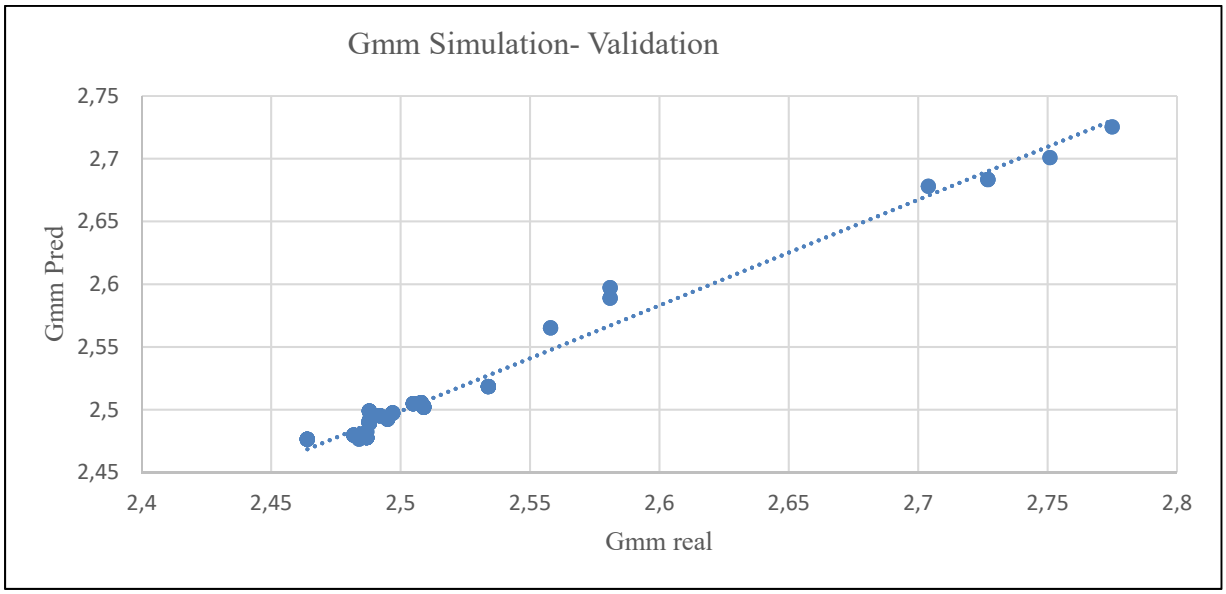


Figure 3. 25 Graphic of the correlation between real and predicted values of Gmm using unseen data by the model 8

Figure 3.26 presents the prediction behaviour during the validation testing epochs from 1 to 29, showing the continuity and precision of the Gmm prediction values.

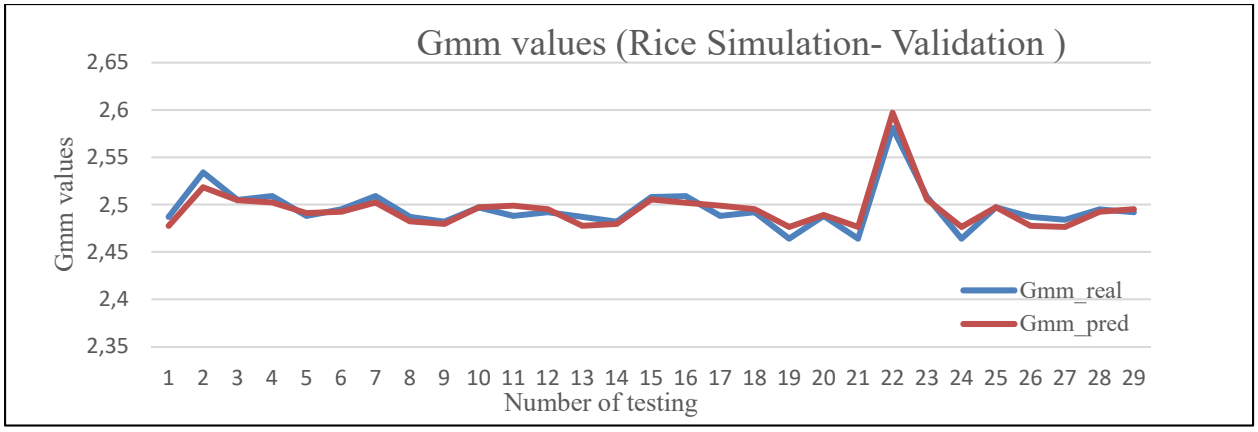


Figure 3. 26 Illustration of correlation between real and predicted values of Gmm of the model 8 during the validation process

3.6 Correlation Analysis for Gmm Predicted

A correlation analysis was achieved to evaluate the impact of aggregates and asphalt binder properties selected as inputs into the Gmb and Gmm prediction models. However, for the model 6-Gmb correlation analysis was avoided based on R^2 equal to 0.4573 as shown in table 3.8. This correlation study was performed for only Gmm predictions, as a result of R^2 of 0.9385 obtained from the validation of model 8 (Gmm) as shown in table 3.9. Additionally, validation testing was performed on models considered in the correlation analysis as shown in table 3.10. This table 3.10 provides summary statistics for the correlation variables used in the analysis. It can be noted in the summary the following characteristics:

1) Before validation with the original model 8
   a) The best performance with the highest R^2 and lowest errors was the model which was trained with only TG as gradation factors (5 inputs). The R^2 for this model was 0.9898 and errors were 0.0059 for MAE and 0.0087 for RMSE;
   b) The four models in which only 1 gradation factor (5 inputs) was used, had similar values regarding R^2 between 0.9898 for TG and 0.9874 for FP; errors were between 0.0059 for TG and 0.0061-0.0067 for GI and G/S concerning MAE; also 0.0087 for TG and 0.0095 for GI regarding RMSE.
   c) When Gsb was removed, the R^2 had the lowest values equal to 0.8884 and the highest error metrics MAE equal to 0.0156, RMSE equivalent to 0.0275;
   d) Concerning binder properties when Gb was removed R^2 was equal to 0.9813. This value is lower than 0.9877 et 0.9886 for R^2 of Pb and kinematic viscosity at 275 F, also errors were higher with MAE equal to 0.0083 and RMSE equivalent to 0.0115 while similar values were obtained for Viscosity and Pb with MAE between (0.064 – 0.060), and RMSE (0.0090-0.0097).

2) After Validation testing with new data
   a) The best performance after validation was the original Model 8 selected which contains 8 inputs with R^2 equal to 0.9385; MAE equal to 0.0080 an RMSE equal to 0.0130
   b) When Gsb was removed as can be seen in Figures 3.26 and 3.27 :
      - ✓ There was no correlation with R^2 equal to 0.0325 after validation testing.
      - ✓ Errors increased MAE from 0.0156 to 0.0519 and RMSE from 0.0275 to 0.0735
   c) When only one gradation factor was used as input:
      - ✓ The model which used GI presented the best performance regarding the 4 gradation factors in which R^2 equal to 0.8685, MAE equal to 0.0095 and RMSE equivalent to 0.0181;
      - ✓ Following by Fp with R^2 0.8484;
      - ✓ G/S with R^2 0.8329 and;
      - ✓ The model that used TG with the lowest R^2 0.8073.
   d) When Pb was removed :
      - ✓ R^2 was equal to 0.8172. This value was the second-lowest R^2. This value decreases from 0.9385 obtained by the validation model 8 with all inputs to 0.8172 without Pb;
      - ✓ Error increased compared with Model 8; MAE from 0.0080 to 0.0129 and RMSE from 0.0130 to 0.205
   e) When Viscosity at 275 F and Gb were removed:
      - ✓ R^2 equal to 0.9156 was found when viscosity was removed while R^2 equivalent to 0.9221 was obtained when Gb was removed;
      - ✓ When the viscosity was removed errors were lower than errors when Gb was removed. The MAE equivalent to 0.0154 and RMSE equal to 0.0220 for the model in which viscosity was removed. MAE equivalent to 0.0147 and RMSE equal to 0.0175 when Gb was removed as input.

Table 3. 10 Performance metrics (before and after the validation process) of models developed in which inputs were changed to evaluate the impact on Gmm predictions

Correlation Analysis (Rice simulation test)						Validation data (166 New pairs)		
Number of inputs	Inputs removed	Outputs	MAE	RMSE	R2	MAE	RMSE	R2
8	-	Gmm-Gse	0.0065	0.0090	0.9887	0.0080	0.0130	0.9385
5 (Only TG)	GI, G/S, Fp		0.0059	0.0087	0.9898	0.0106	0.0203	0.8073
5 (Only GI)	TG, G/S, Fp		0.0067	0.0095	0.9874	0.0095	0.0181	0.8685
5 (Only G/S)	GI, TG, Fp		0.0059	0.0088	0.9894	0.0095	0.0196	0.8329
5 (Only Fp)	GI, TG, G/S		0.0061	0.0090	0.9890	0.0100	0.0185	0.8484
7	Gsb		0.0156	0.0275	0.8884	**0.0519**	**0.0735**	**0.0325**
7	Visc 275		0.0064	0.0090	0.9886	0.0154	0.0220	0.9156
7	Gb		0.0083	0.0115	0.9813	0.0147	0.0175	0.9221
7	Pb		0.0060	0.0097	0.9877	0.0129	0.0205	0.8172

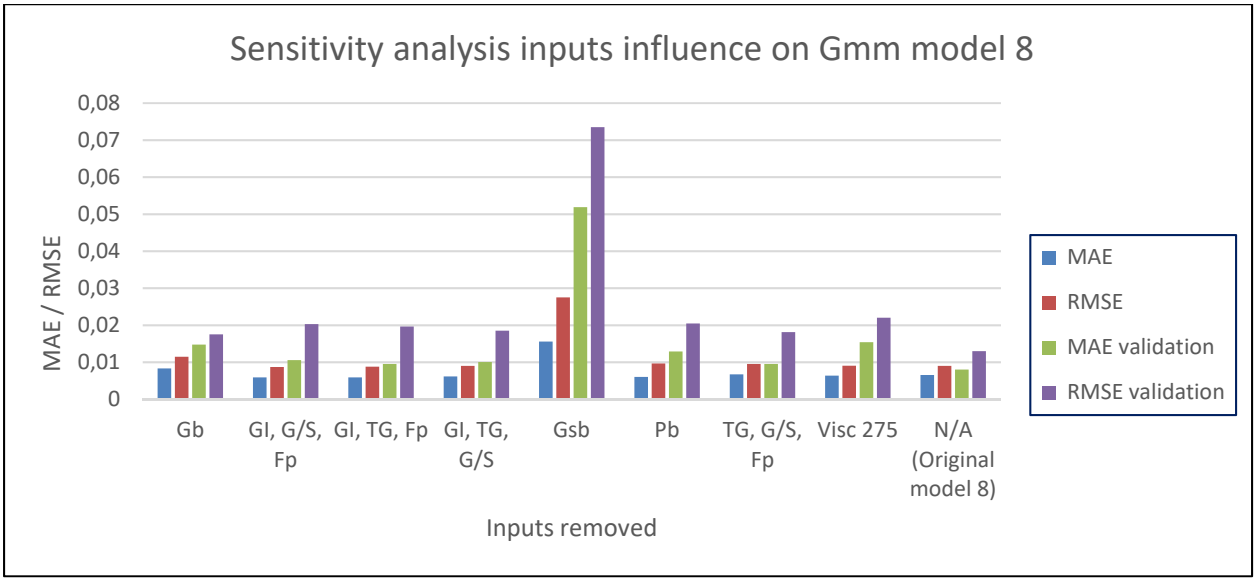


Figure 3. 27 Illustration of input the influence on the performance metrics (MAE, RMSE) before and after validation of the original model 8 to predict Gmm values.

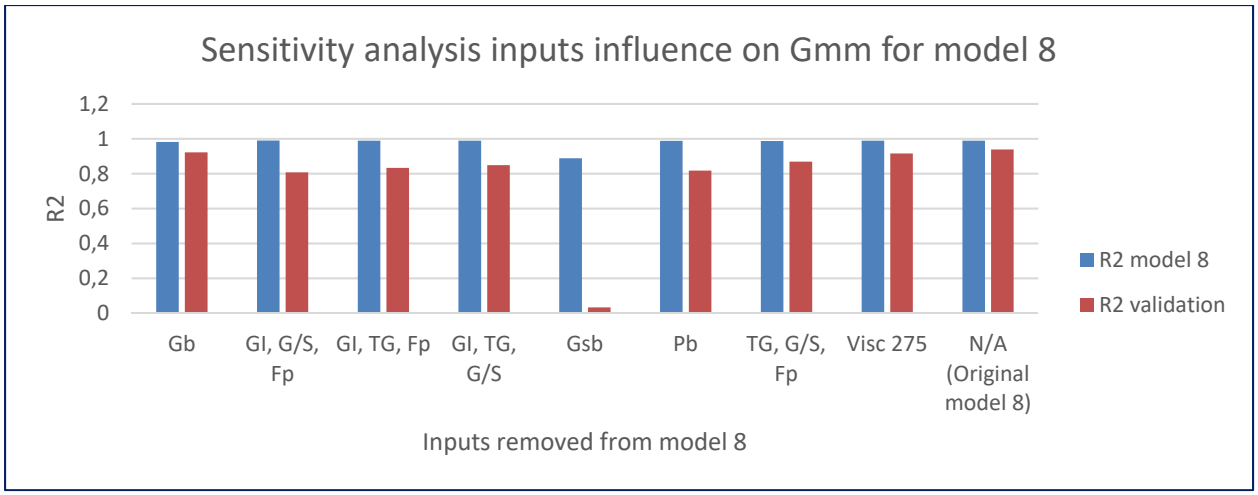


Figure 3. 28 Illustration of inputs the influence on the correlation coefficient R^2 before and after validation of the original model 8 to predict Gmm values

3.7 Discussion

3.7.1 Assessing the model selection process

It was observed that the SGC phenomenon is more complex than the Rice test. In fact, when models for the rice test and the SGC were trained under the same DNN architecture (layers and nodes), thus overfitting condition was evidence for the rice test training. In effect, Rice test model (1 and 2) trained with only 3 hidden layers and 150 nodes were found overfitted as can be seen in Figures 3.8 and 3.9. However, SGC models trained with 4, 5, and 6 layers and 150, 200, 350, 450 nodes, were found well fitted as shown in Figures 3.5, 3.6, 3.7, and in annexe I. This finding is in line with the hypothesis of the phenomenon of the Rice test is simpler than the SGC phenomenon in which SGC could need more data to effectively be generalized. Typically, when the number of layers and nodes is increased into the DNN, thus expanding the learning capacity for the neural network learns from the data developing the non-linear connections. However, the risk of developing a model that can memorize the dataset increase, thus incapacitating the model to adequately predict when new data is used. Moreover, as can be seen in Figures 3.10, 3.11, 3.12, 3.13, 3.14 and in annexe II; a better performance correlation coefficient R2 was obtained when models for Rice test were trained with 2 outputs

as shown in table 3.5. This behaviour could be attributed to the effect that the second output (Gse) produced in the backpropagation of error in which connection weights are adjusted with the intention of aid reconcile the differences between the real Gmm values and predicted Gmm outcomes for subsequent forward passes into the Neural network interaction. In effect, the idea of adding the Gse as output was optimizing the model by adjusting the better the connection weights, despite the fact that Gse performance was completely overfitted for Model 8 selected. In addition, Gse was found overfitted for all models after the validation process. This overfitting found in Gse prediction is not considered as a limitation of this research. To clarify, the main objective of this work was to accurately predict Gmm values simulating the Rice test because it is a common practice obtaining Gse by simple calculation after Rice test is run to obtain Gmm.

Moreover, the first performance results obtained from both selected models developed (Model 6-Gbm and Model 8-Gmm) based on MAE, RMSE as and R2 as can be seen in tables 3.6 and 3.7, show that these models accurately predict the outputs (Gmb and Gmm). However, an external validity using unknown data was essential for avoiding working on correlation analysis of an overfitted model, thereby drawing wrong conclusions regarding the influence of independent variables into the phenomenon.

This external validity of both models (Model 6 and Model 8) revealed that the former, model 6 (Gmb-prediction) was overfitted with R2 equal to 0.4385 as shown in table 3.8 and Figure 3.21 (dispersion points). Thus, more data would be required to solve the problem and adequately generalize the phenomenon of SGC. Additionally, the disparity observed in Figure 3.21, between the 2 curves reveals that the model could not adequately predict Gmb values with new data. This result supports the previous assumption concerning the SGC model would need more data than the Rice test simulation. Conversely, the latter, Model 8-Gmm was found well fitted. This model had R2 equal to 0.9385 after testing validation with new data as can be seen in table 3.9. It confirms that the Rice test phenomenon apparently was generalized by the Model 8 developed.

3.7.2 RMSE and standard deviation of acceptance (AASTHO T209 and ASTM D2041)

Although the RMSE is a quantification of the average difference of observations from a real value (Bias). This could be compared with the SD which is a quantification of the average difference of observations from the mean value of the samples (precision and accuracy). Then, little RMSE values indicate that the mean of Gmm prediction values is close to the mean of the Gmm true values, which suggests that the SD is close to RMSE (Meyer, 2012). Thus, accuracy and precision were achieved for the model. In fact, both original models (model 6-Gmb and model 8-Gmm) before validation presented little RMSE values and the SDs of the prediction real values were similar.

Furthermore, the Model 8 (Gmm-prediction) validation results could indicate that the model remained with the accuracy and precision due to both standard deviations were closed in which the SD was 0.0523 for prediction values and SD real equal to 0.0613 as shown in Figure 3.24 and RMSE 0.0130 as shown in table 3.9. However, model 6 (Gmb-prediction) was found overfitted after validation testing in which RMSE increased from 0.0127 to 0.1004 as shown in tables 3.6 and 3.8. In addition, a large difference between the SD for predictions equivalent to 0.055 and SD for real values equal to 0.1217 as shown in Figure 3.21. Thereby, model 6 was found with precision, nonetheless inaccurate (Biased).

In particular, it is known that a constant value of Gmb, changes in Gmm by +0.01, thereby changing Va by +0.4% (FHWA, 2010). This concept permits establish that errors RMSE equivalent to 0.0090 for model 8 (Gmm-Predictions) would represent changes in Va by 0.36%. However, considering the results after validation for model 8 with RMSE equivalent to 0.0130 would change Va by 0.52%. Moreover, the Rice tests (AASHTO T209 and ASTM D2041) suggest multilaboratory precision values. These SDs values for multilaboratory are 0.0064 and 0.016 regarding no supplemental procedures (dryback) for absorptive aggregate mixtures (FHWA, 2010). These precision values can be compared with the model 8 errors in which the SD equal to 0.0523 for prediction values is higher than the acceptable limits. Thus, the Gmm predictions of model 8 seem to dissatisfy with the AASHTO and ASTM standards. However, following the common practice that duplicating specific gravity results should be repeated only

if they differ more than 0.02 when it has been run by the same operator. Thus, model 8 with RMSE equal to 0.0130 would perform with this traditional practice in the lab. In effect, model 8 showed low precision after validation with SD equal to 0.0523; however, RMSE equal to 0.0130 seems to be accurate (unbiased) despite the fact that it would represent changes in Va by 0.52% as mentioned in the literature by (FHWA, 2010).

3.7.3 Possible causes of overfitting Model 6-Gmb

Model 6 (Gmb-prediction) was found in overfitting probably because:

1) The dataset used contains similar pairs (homogeneous) and more data from other clusters or groups is essential to provide more information into the neural network. In fact, data-driven models must have a diversity dataset which represents as much as possible the phenomenon to be modelled. Moreover, when a data set is focused on only one section of this spectrum as the case of Model 6, models tend to be responsive only on a small part of the spectrum.
2) SGC phenomenon is more complex than the Rice test and only 26,620 pairs were used to train the algorithm. This quantity of pairs (26,260) represents approximately 8.5% of the pairs used in the Rice test prediction (model 8). In fact, even with the same quantity of pairs for both models (6 and 8) the probability to develop an overfitting model for The SGC is high.

The complexity of the SGC phenomenon probably needs more information from other inputs variables regarding the aggregates like absorption, angularity, and surface texture should be added to the neural network. It seems that these aggregates properties influence in the phenomenon, thus making the SGC phenomenon more complex to simulate under the same inputs. These input variables available and used in the training of the model 6 were limited regarding aggregate properties information.

However, despite most of the values in the unseen dataset are into the training range values as can be seen in tables 3.1 and 3.3, it seems that this model was probably avoided and judged wrongly overfitted when the unseen data for validation includes values out of the range in which the model was training. The table 3.1 show that the model was trained between 7 and149

gyrations and table 3.3 shows that the unknown data set has a range of values from 6 to 205 gyration. This variation can be seen in figure 3.29.

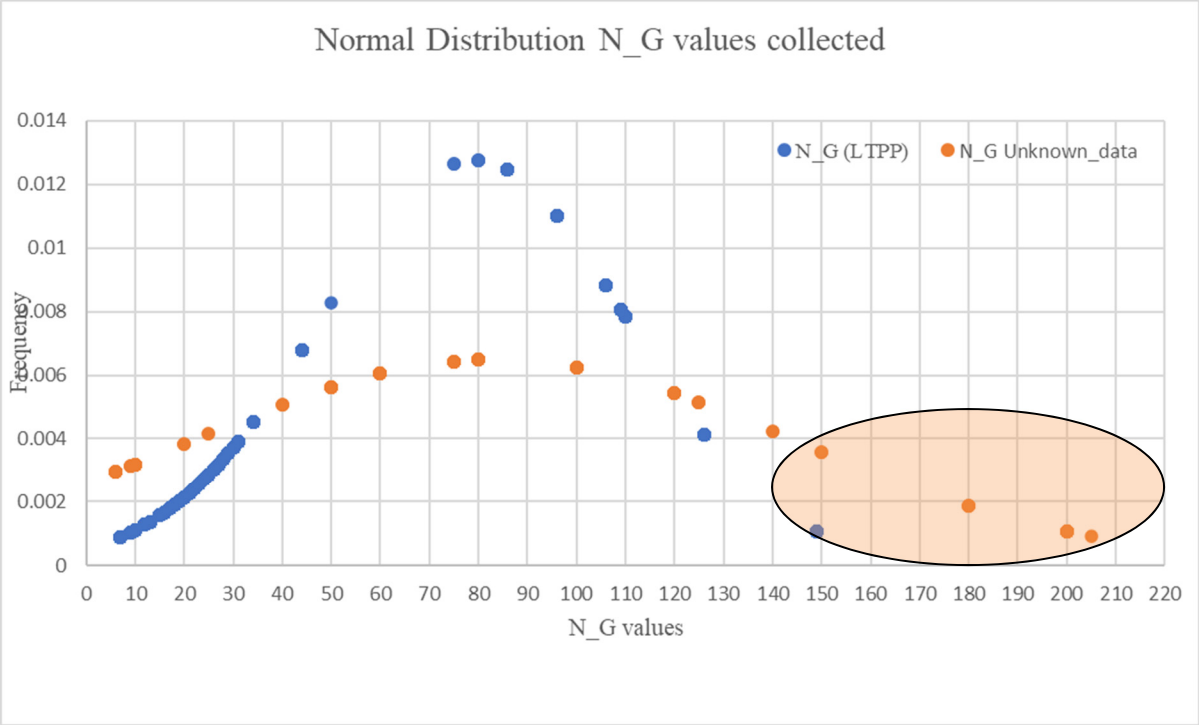


Figure 3. 29 Distribution of data collected from LTPP and unseen data for N_G values

Additionally, the unknown data set has Gmb values out of the training range as can be seen in the tables 3.1 and 3.3 in which the training has values between 2.229 and 2.644 and the range values of the unseen dataset is between 2.016 and 2.662 as shown if figure 3.30.

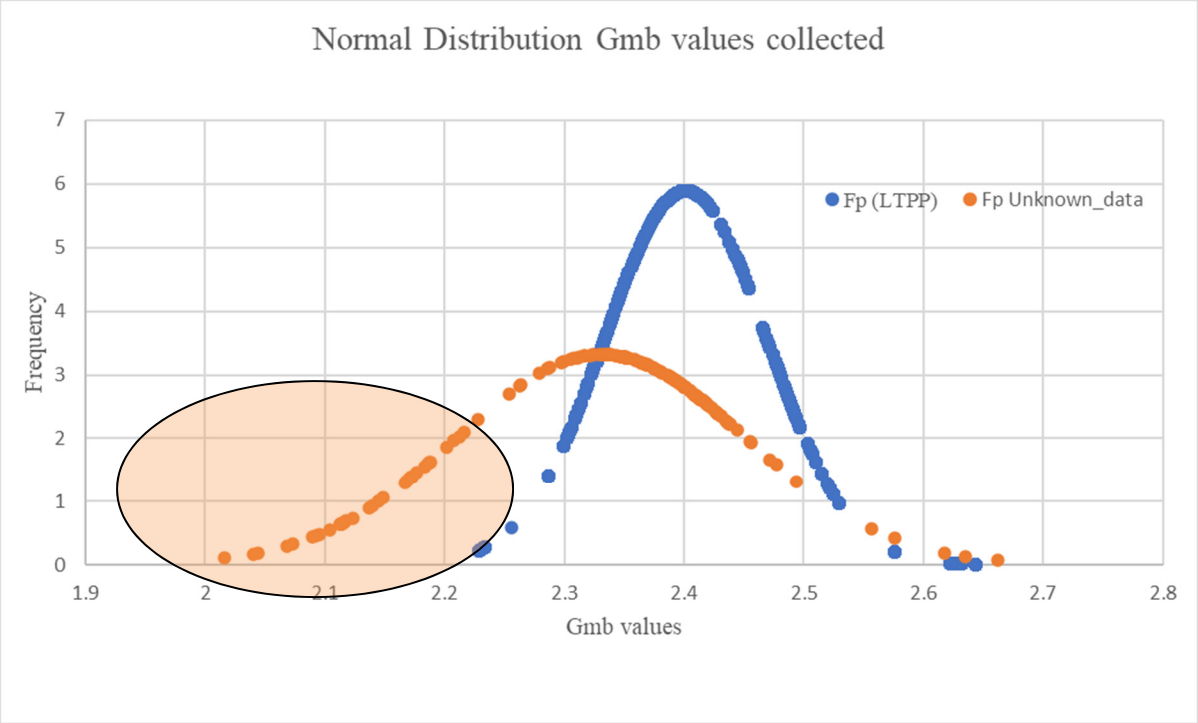


Figure 3. 30 Distribution of data collected from LTPP and unseen data for Gmb values

It is known the limitation of the ANN algorithm to extrapolate results. For this reason, a new validation should be done using the unseen data set by eliminating the pairs values which are out of the training range.

3.7.4 Correlation Analysis Rice Test (Gmm-Prediction Model)

Correlation analysis should be avoided when the phenomenon generalization of the model is unclear, thus drawing wrong conclusions due to the model performs the same with different inputs. This approach was considered to avoid a correlation study of model 6 in which Gmb prediction was performed. Only a correlation analysis was considered for model 8 regarding Gmm prediction model.

3.7.4.1 Regarding Gradation

The original model 8 showed that all gradation factors (TG, GI, G/S, and FP) had similar performance regarding R^2. These results could be explained by the fact that the Gmm is a volumetric property and because no voids are present in the rice test to obtain Gmm which exposes that gradation does not provide significant information into the DNN for the Gmm prediction. Also, this output Gmm is not a direct measure of HMA performance like rutting, fatigue, resistance (MR), workability, and permeability which were correlated with these factors in the literature. Thus, it could be concluded that gradation factors provide no influence on Gmm. Furthermore, by using the 4-gradation factor into the neural network seems to increase the noise, thereby reducing performance R^2 as shown in table 3.10 in which R^2 increased from 0.9887 to 0.9890, 0.9894, 0.9898 when 3 factors were removed.

However, after validation testing differences among models regarding R^2 and errors were observed. This indicates that gradation factors had a small effect on output Gmm. In addition, these results reveal that GI is the most correlated gradation factor with Gmm, following by FP in second place, G/S in third place, and TG as the gradation factor with the lowest correlation as can be seen in table 3.10, despite the similar errors were found among all models trained with only one gradation factor. Additionally, when the 4-gradation factors were included in model 8 the coefficient R^2 equal to 0.9385 was the highest after validation testing, thus indicating that adding the 4-input gradation factors provided more information into the neural network than just using one factor. This result also permits to confirm that no noise was produced by using the 4 factors.

3.7.4.2 Regarding Specific Gravity

The original model 8 showed that Gsb provides the most information to the DNN system in which R^2 decreased from 0.9887 to 0.8884 when Gsb was removed as can be seen in table 3.10 and Figure 3.28. This result could be elucidated by noticing that the input Gsb and outputs Gmm are both specific gravities. In effect, the calculation of Gsb contains gradation information of the aggregate mix, thus increasing the information that this input variable

provides into the neural network. Furthermore, the performance metrics after validation testing indicate that Gsb was the most important inputs of the model as shown in Figures 3.27 and 3.28. This was evident when no correlation was found with R^2 equal to 0.0325 and the highest errors as shown in table 3.10. This important role of Gsb for Gmm prediction can be justified logically because around 95% of an asphalt mix is composed by aggregates, thereby showing that predictions provide by the model 8 presented in this research are scientifically sound.
Moreover, this relative importance found regarding the Gsb is consistent with data from the study of (Gong et al., 2019) in which Gsb was found as the input variable with the most impact on alligator cracking, transverse cracking and rutting phenomena, by being the transverse cracking the most sensitive to the Gsb changes. In fact, The Gsb was found more important than the service age of section and the indirect tensile strength (IDT) for the transverse cracking phenomenon. Furthermore, Gsb was found the most significant input to predict Gmm in the recent study developed by (Dalhat & Osman, 2022) in which the database of LTPP was used. However, the amount of data employed in the present study is almost 75 times higher regarding the quantity of data used by (Dalhat & Osman, 2022),

3.7.4.3 Regarding Asphalt Binder Properties

Model 8 before validation testing presented a reduction of R^2 from 0.9887 to 0.9877 when Pb was removed as can be seen in table 3.10. This result would permit elucidating that Pb influences the Gmm prediction. However, this reduction is lower than the reduction of R^2 from 0.9887 to 0.9813 produced when Gb was removed. This finding reveals that the specific gravity of the binder gives more information into the neural network than Pb. However, these results do not fit with the theory because it is known that changes in the Pb produce direct changes in the percentage of aggregates (Ps), in addition, no compaction (air voids) is involved in the Rice test to obtain the Gmm. Thus, the volume of binder, which is more than twice the volume of aggregates, produces sensitive changes into Gmm when binder is added. In fact, this finding could draw wrong conclusion considering the Gb more important than Pb in the Gmm

prediction phenomenon. In effect, this illogical finding could indicate that the model is overfitted. This probably overfitting shows that the model, in which Pb was removed, has memorized the dataset being Pb unimportant into the DNN to predict Gmm of the same dataset with a high correlation coefficient R^2. To clarify this issue the validation testing applied to the models of correlation study provided important results that differ from the original correlation analysis. As can be seen in table 3.10 the model in which Pb was removed the coefficient R^2 decreased from 0.9385 to 0.8172. On the other hand, in the model in which Gb was removed the coefficient R^2 was 0.9221. This outcome coincides with the theory of the effect of binder content (Pb) on Gmm in which Pb has more impact on Gmm than Gb.

Moreover, the coefficient R^2 was found less sensitive to the binder properties like Gb and viscosity at 275 ºF than Pb for validation testing. These coefficients R^2 were 0.9221 for Gb and 0.9156 for the kinematic viscosity at 275 ºF. This result indicates that these variables were redundant into the neural network. Thus, this model would perform similar regardless of Gb or Viscosity is used. In fact, despite the variability of the data collected for the kinematic viscosity at 275 ºF viscosity in which 95% of the values are between 284 and 1465 cSt with a SD of 590.57 as shown in Figure 3.32; in contrast to the variability of Gb values collected in which these values are between 1.015 and 1.04 t/m3 with a SD of 0.0116 as can be seen in Figure 3.31.

These results suggest that the kinematic viscosity at 275 ºF does not play an important role in Gmm predictions when original binder (no modified) is used. This supposition of no modified binder was used in the training model process might be reasonable because as shown in Figure 3.30, this input viscosity data fits with the range from 30 to 6000 mm2/s [cSt] that the ASTM D2170/D2170-18 allows for precision. Moreover, this method is based on the assumption that asphalt binders behave as Newtonian fluids. It is clearly indicated that modified asphalt binders are typically non-Newtonian under the conditions of this kinematic viscosity test. Thus, invalid shear rates can be achieved when the flow is non-Newtonian in a capillary tube and verification with viscometers having different-sized capillary tubes should be performed.

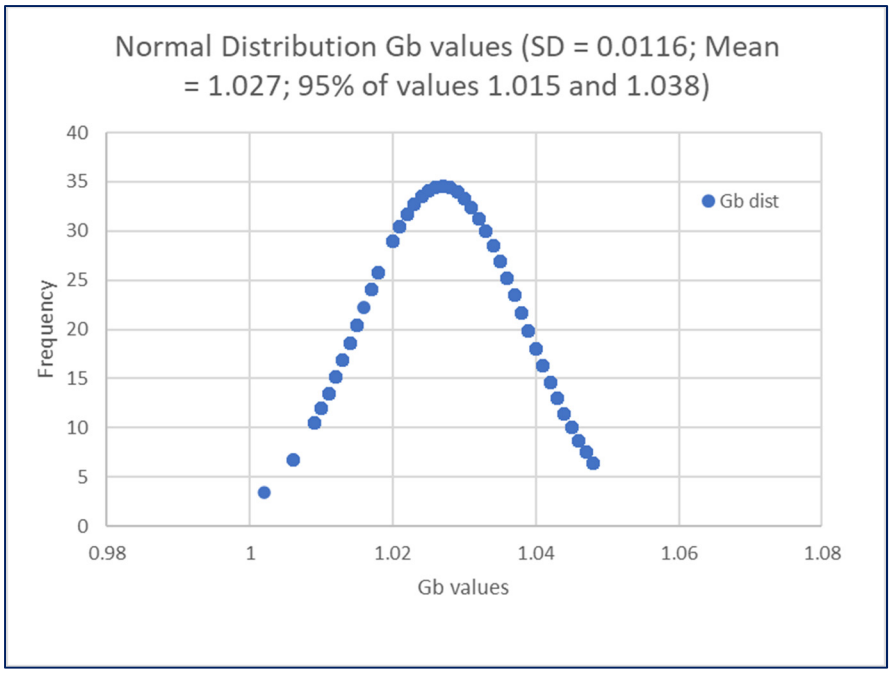


Figure 3. 31 Normal distribution curve of input data of Gb values

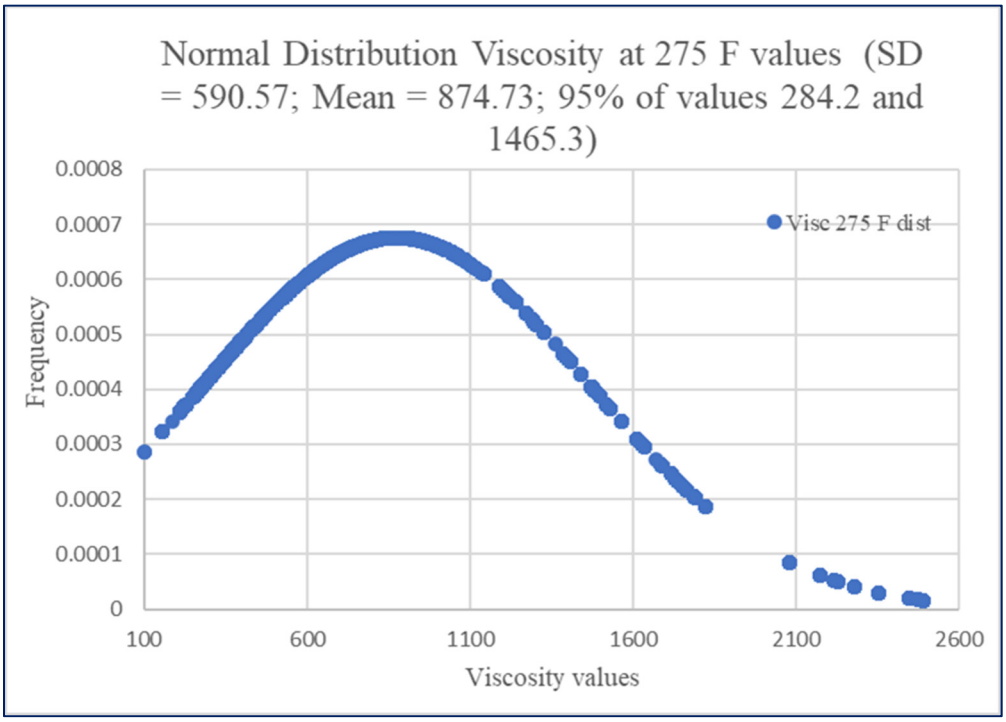


Figure 3. 32 Normal distribution curve of input data of viscosity values

CONCLUSION

This study aimed at developing predictive models of HMA volumetric parameters, mainly Gmb (from SGC method) and the maximum specific gravity Gmm (from Rice test), using Deep Neural Network (DNN). The considered inputs were four aggregate packing quantification factors (FP, G/S, GI, TG), one aggregate volumetric property (Gsb) and three binder properties (Gb, Pb, viscosity at 275 F). This work was split in two stages: model selection and model validation, with distinct datasets. Also, it was reviewed the state-of-the-art concepts of each component of models proposed by showing how each part contributes to achieving an optimization of the volumetric HMA design to predict Gmb and Gmm values using deep neural networks (DNN) under supervised machine learning technique. The overall findings can be listed as below.

1) Reasonable results can be obtained using machine learning techniques to predict HMA volumetric properties by evaluating performance model generalizations through learning curves.
2) The coefficient of correlation of R^2 is not enough to assess predictions performance by which DNN models are developed because this metric is unclear regarding generalizations.
3) Validation process using a dataset unseen by the model is essential to avoid overfitting problems.
4) The phenomenon of the Rice test needs less DNN complexity compared to the SGC phenomenon in which SGC could need more data to effectively be generalized. This necessity of more data can be covered increasing the amount of data using the same input variables or adding more input variables. This was confirmed when external validity (validation process) was applied for both models (Model 6 (Gmb-prediction) and Model 8 (Gmm-prediction)), thereby revealing that the model 6 was overfitted.
5) The second output of Gmm predictive model, The Effective specific gravity of aggregate (Gse) added into the DNN to predict Gmm values, produced in the backpropagation of error better connection weights, thereby optimizing the model was found despite the fact that Gse performance was completely overfitted.

6) As expected from a physical standpoint, bulk specific gravity for the total aggregate (Gsb) provides the most information to the DNN system regarding the Rice test phenomenon to predict Gmm.
7) The binder properties into the Gmm prediction model after validation indicates:
   a) Asphalt content (Pb) has more impact on Gmm than Specific gravity of asphalt (Gb)
   b) The Gb and the kinematic viscosity at 275 ^{o}F are redundant into the neural network, despite the variability of the data collected for the kinematic viscosity at 275 oF. Thus, the model performs similar regardless of Gb or Viscosity is used as input.
8) The model 8 (Gmm-prediction) showed low precision after validation; however, it seems to be accurate despite the fact that it would represent high changes in Va by 0.52% according to the literature (FHWA, 2010).
9) The Rice test model to predict Gmm of asphalt mixtures has been evaluated by using a DNN model. The primary variable which was investigated is the aggregate gradation. Indeed, the sensitivity analysis between input parameters and output showed that some of the gradation factors were plausible correlated with the Gmm. However, all the 4 gradation factors were useful, as they reduced noise in the model 8 (Gmm-prediction). Moreover, GI is the most correlated gradation factor with Gmm, following by Fp in second place, G/S in third place, and TG as the gradation factor with the lowest correlation.

This work emphasizes and corroborates the importance of the generalization evaluation by displaying learning curves and the validation using unseen data by the models previously developed. Further studies are suggested to develop bulk specific gravity of the HMA (Gmb) prediction models, thus completing the main part of the HMA volumetric design using artificial intelligent. This can be developed by considering other or/and more variables to be added into the neural network like absorption, angularity, and surface texture.

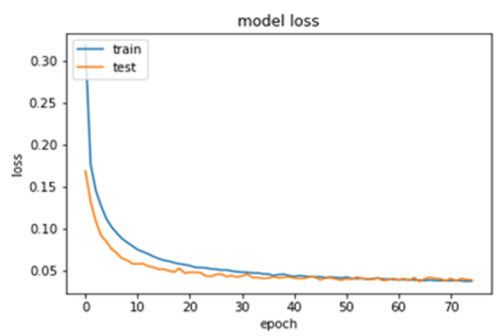
This work is an initial stage of achieving the target of using AI in order to limit the lab work and make the process more efficient. Then, developing models and confirming that AI is applicable in HMA when enough data is accessible was essential. For this reason, beginning with an uncomplicated phenomenon like Gmm allowed the author to achieve one of the main benefits of this research which is to understand how the MLT specially ANN can be applied

in the field of HMA to properly obtain accurate predictions. In the meantime, the constrains regarding the amount of data necessary to adequately generalize the phenomenon can be overcome by applying the transfer learning technique where performance algorithms can be trained with more easily collected data. The result of evaluation of this complex tool (ANN) in HMA will help to reach the modification to be able to adequately predict behaviours. In fact, this work suggests that the proposed model 8 structure and weights could be used as a starting point to predict other mixture properties (e.g. rutting, fatigue, permeability, workability and resilient module) in which gradation could have a more important role compared to the role in Gmm-prediction model. Additionally, the MLT random forest algorithm could be applied to this data to estimate the impact of asphalt mixtures components on the volumetric properties made in this work.

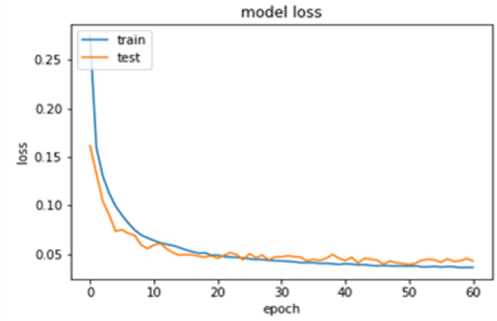
ANNEX I

LEARNING CURVES OF MODELS 2, 3 AND 5 FOR SGC SIMULATION

DNN Architecture of the Model 2 for SGC simulation						
Training (Model 2)	Outputs (Gmb)	% training (75)	Layers (4)	Nodes (200)	Dropout (0.15)	R2 (0.9608)



DNN Architecture of the Model 3 for SGC simulation						
Training (Model 3)	Outputs (Gmb)	% training (75)	Layers (5)	Nodes (350)	Dropout (0.15)	R2 (0.9533)



DNN Architecture of the Model 5 for SGC simulation						
Training (Model 5)	Outputs (Gmb)	% training (70)	Layers (6)	Nodes (450)	Dropout (0.10)	R2 (0.9586)

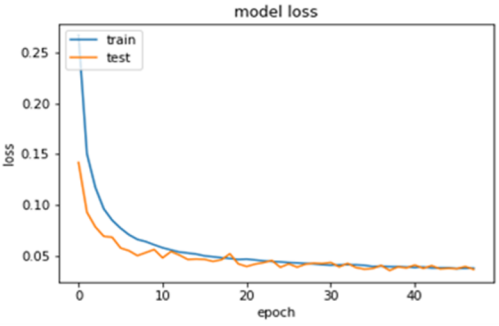
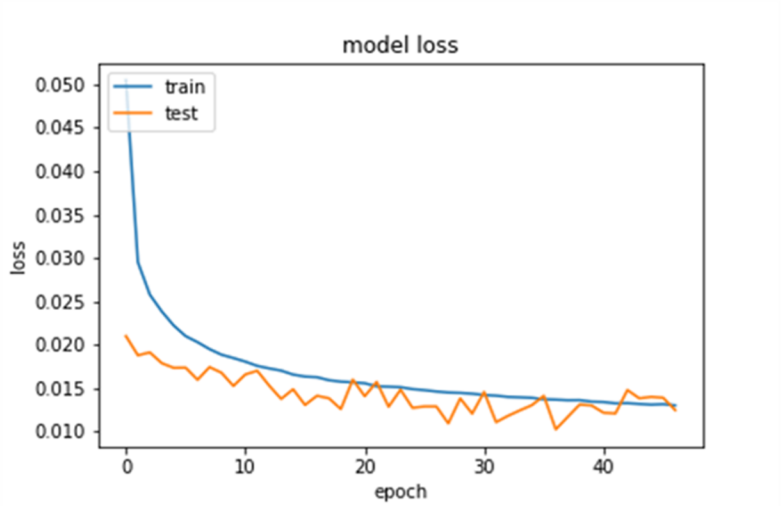


Figure A. 1 Learning curves of models 2, 3 and 5 to predict Gmb values. The curve (orange) of test loss decreases steadily and then stabilizes after epoch 20. The training loss curve (blue)

ANNEX II

LEARNING CURVES OF MODELS 4 AND 6 FOR RICE TEST SIMULATION

DNN Architecture of the Model 4 for Rice test simulation						
Training (Model 4)	Outputs (Gmm)	% training (60)	Layers (2)	Nodes (150)	Dropout (0.15)	R2 (0.9873)



DNN Architecture of the Model 6 for Rice test simulation						
Training (Model 6)	Outputs (Gmm)	% training (70)	Layers (2)	Nodes (150)	Dropout (0.15)	R2 (0.9869)

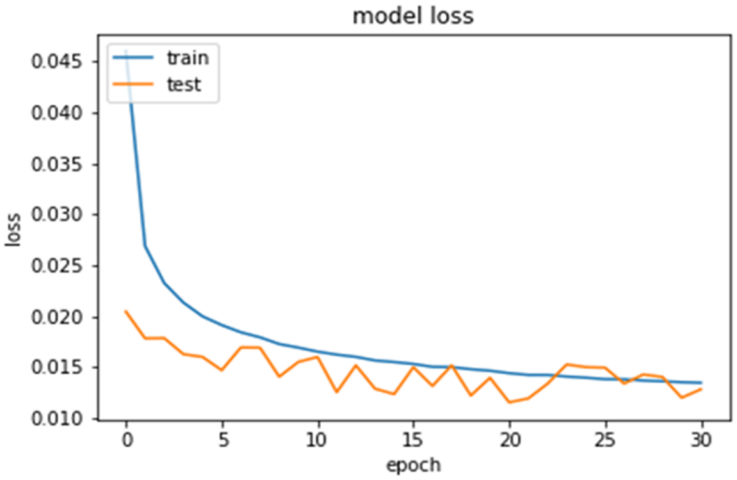


Figure A. 2 Learning curves of model 4 and 6 to predict Gmm values. The curve (orange) of test loss decreases steadily and then stabilizes after epoch 20. The training loss curve (blue

BIBLIOGRAPHY

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. *Heliyon, 4*(11), e00938.

Al-Mosawe, H. (2016). *Prediction of permanent deformation in asphalt mixtures.* University of Nottingham,

Al-Mosawe, H., & Thom, N. (2016). *Prediction of permanent deformation in asphalt mixtures.* University of Nottingham,

Anyoha, R. (2017). The History of Artificial Intelligence. Retrieved from http://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/

Arif, A. (2020). CROSS VALIDATION IN MACHINE LEARNING. Retrieved from https://dataaspirant.com/cross-validation/

Asphalt-Institute. (2001). *Superpave Mix Design (SP-2)* (Vol. Third Edition).

Asphalt-Institute. (2014). *MS-2 Asphalt Mix Design Methods* (7 th (Seventh) ed.).

Aurilio, V., Pine, W., & Lum, P. (2005). *The Bailey Method Achieving Volumetrics and HMA Compactability.* Paper presented at the Proceedings of The Annual Conference-Canadian Technical Asphalt Association.

Benko, A., & Lányi, C. S. (2009). History of artificial intelligence. In *Encyclopedia of Information Science and Technology, Second Edition* (pp. 1759-1762): IGI Global.

Brownlee, J. (2019). *Better Deep LearningMachine Learning Mastery* (v1.5 ed.).

Buchanan, B. G. (2005). A (very) brief history of artificial intelligence. *Ai Magazine, 26*(4), 53.

Ceylan, H., Kim, S., & Gopalakrishnan, K. (2007). Hot mix asphalt dynamic modulus prediction models using neural networks approach.

Cooley Jr, L. A., Brown, E. R., & Maghsoodloo, S. (2001). Development of critical field permeability and pavement density values for coarse-graded Superpave pavements. *Journal of the Transportation Research Board*(1761).

Dalhat, M. A., & Osman, S. A. (2022). Artificial Neural Network Modeling of Theoretical Maximum Specific Gravity for Asphalt Concrete Mix. *International Journal of Pavement Research and Technology*. doi:10.1007/s42947-022-00244-0

Fadhil, T. H., Ahmed, T. M., & Al Mashhadany, Y. I. (2022). Application of Artificial Neural Networks as Design Tool for Hot Mix Asphalt. *International Journal of Pavement Research and Technology, 15*(2), 269-283. doi:10.1007/s42947-021-00065-7

Feipeng Xiao, P. D., P.E1; and Serji N. Amirkhanian, Ph.D.2. (2009). Artificial Neural Network Approach to Estimating Stiffness Behavior of Rubberized Asphalt Concrete Containing Reclaimed Asphalt Pavement. *Journal of Transportation Engineering, 135*. doi:10.1061/共ASCE夫TE.1943-5436.0000014

FHWA. (2010). *A REVIEW OF AGGREGATE AND ASPHALT MIXTURE SPECIFIC GRAVITY MEASUREMENTS AND THEIR IMPACTS ON ASPHALT MIX DESIGN PROPERTIES AND MIX ACCEPTANCE* Retrieved from https://www.fhwa.dot.gov/pavement/materials/pubs/hif11033/hif11033.pdf

FHWA. (2019). Long Term Pavement Performance (LTPP). Retrieved from https://infopave.fhwa.dot.gov/Data/DataSelection

Gong, H., Sun, Y., Hu, W., Polaczyk, P. A., & Huang, B. (2019). Investigating impacts of asphalt mixture properties on pavement performance using LTPP data through random forests. *Construction and Building Materials, 204*, 203-212.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning (Adaptive Computation and Machine Learning series)*: MIT Press.

Groen, I. I., Silson, E. H., & Baker, C. I. (2017). Contributions of low-and high-level properties to neural processing of visual scenes in the human brain. *Philosophical Transactions of the Royal Society B: Biological Sciences, 372*(1714), 20160102.

Guarin, A. (2009). *Interstitial component characterization to evaluate asphalt mixture performance.* (DOCTOR OF PHILOSOPHY). University of Florida, Florida.

Haykin, S. S. (2009). Neural networks and learning machines/Simon Haykin. In: New York: Prentice Hall.

Heaton, J. (2013). *AIFH, Volume 1: Fundamental Algorithms*.

Henderson, R., Herrington, P., Patrick, J., Kathirgamanathan, P., & Cook, S. (2011). Analysis of Particle Orientation in Compacted Unbound Aggregate. *Road Materials and Pavement Design, 12*(1), 115-127. doi:10.1080/14680629.2011.9690355

Huyan, J. (2019). *Development of machine learning based analytical tools for pavement performance assessment and crack detection.* University of Waterloo,

Isola, M. (2014). *Effect of interstitial volume characteristics on asphalt mixture durability and cracking performance.* (DOCTOR OF PHILOSOPHY). University of Florida,

Kim, S. (2006). *Identification and Assessment of the Dominant Aggregate Size Range of Asphalt Mixture.* (DOCTOR OF PHILOSOPHY). University of Florida, Florida.

Liu, Y., Zhao, T., Ju, W., & Shi, S. (2017). Materials discovery and design using machine learning. *Journal of Materiomics, 3*(3), 159-177. doi:https://doi.org/10.1016/j.jmat.2017.08.002

Maind, S. B., & Wankar, P. (2014). Research paper on basic of artificial neural network. *International Journal on Recent and Innovation Trends in Computing and Communication, 2*(1), 96-100.

Masad, E., Tashman, L., Somedavan, N., & Little, D. (2002). Micromechanics-based analysis of stiffness anisotropy in asphalt mixtures. *Journal of Materials in Civil Engineering, 14*(5), 374-383.

Meyer, T. (2012). Root Mean Square Error Compared to, and Contrasted with, Standard Deviation. *Surveying and Land Information Science, 72*.

Mo, W., Huang, Y., Zhang, S., Ip, E., Kilper, D. C., Aono, Y., & Tajima, T. (2018, 11-15 March 2018). *ANN-Based Transfer Learning for QoT Prediction in Real-Time Mixed Line-Rate Systems.* Paper presented at the 2018 Optical Fiber Communications Conference and Exposition (OFC).

Ozgan, E. (2011). Artificial neural network based modelling of the Marshall Stability of asphalt concrete. *Expert Systems with Applications, 38*(5), 6025-6030. doi:10.1016/j.eswa.2010.11.018

Ozturk, H. I., & Kutay, M. E. (2014). An artificial neural network model for virtual Superpave asphalt mixture design. *International Journal of Pavement Engineering, 15*(2), 151-162. doi:10.1080/10298436.2013.808341

Perraton, D., Meunier, M., & Carter, A. (2007). Application of granular packing methods to the mix design of Stone Matrix Asphalts (SMA). *Bulletin de Liaison des Ponts et Chaussées*(270-271).

Quebec, T. (2019). Tome VII Materiaux Chapter 4 Norme 4202. Enrobés à chaud formulés selon la méthode de formulation du Laboratoire des chaussées

In T. Quebec (Ed.).

Russell, S. J. (2010). *Artificial intelligence : a modern approach*: Third edition. Upper Saddle River, N.J. : Prentice Hall, [2010] ©2010.

Sánchez-Leal, F. (2003). *RAMCODES: A new approach to soil compaction quality control.* Paper presented at the Proc., 12th Panamerican Conf. on Soil Mechanics and Geotechnical Engineering.

Sánchez-Leal, F. J. (2007). Gradation Chart for Asphalt Mixes: Development. *Journal of Materials in Civil Engineering, 19*(2), 185-197. doi:doi:10.1061/(ASCE)0899-1561(2007)19:2(185)

Sanchez-Leal Jucá, J. F. T., De Campos, T. M., & Marinho, F. A. (2002). *Correlation of maximun density and optimum water content with index properties Unsaturated Soils: Proceedings of the Third International Conference on Unsaturated Soils, UNSAT 2002, 10-13 March 2002, Recife, Brazil* (Vol. 2): CRC Press.

Sanchez Leal, F., Garnica Anguas, P., Gómez López, J., & Pérez García, N. (2002). RAMCODES: Metodología racional para el análisis de densificación y resistencia de geomateriales compactados. *Publicación Técnica*(200).

Sebaaly, H., Varma, S., & Maina, J. W. (2018). Optimizing asphalt mix design process using artificial neural network and genetic algorithm. *Construction and Building Materials, 168*, 660-670. doi:https://doi.org/10.1016/j.conbuildmat.2018.02.118

Setiawan, A., Suparma, L., & Mulyono, A. (2016). *The Effect of Aggregate Gradation on Tensile Strength of Asphalt Concrete*.

Setiawan, A., Suparma, L. B., & Mulyono, A. T. (2016). The Effect of Aggregate Gradation on Workability of Asphalt Concrete. *International Journal of Engineering and Technology (IJET), 8*(4), 1750-1757.

Setiawan, A., Suparma, L. B., & Mulyono, A. T. (2017). Modelling Effect of Aggregate Gradation and Bitumen Content on Marshall Properties of Asphalt Concrete. *International Journal on Advanced Science, Engineering and Information Technology, 7*(2), 359-365.

Shen, S., & Yu, H. (2011). Analysis of Aggregate Gradation and Packing for Easy Estimation of Hot-Mix-Asphalt Voids in Mineral Aggregate. *Journal of Materials in Civil Engineering, 23*(5), 664-672. doi:doi:10.1061/(ASCE)MT.1943-5533.0000224

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research, 15*, 1929-1958.

Stanford-University. (2021). CS231n-Convolutional Neural Networks for Visual Recognition. Retrieved from https://cs231n.github.io/neural-networks-3/

Tashman, L. S., Masad, E., Peterson, B., & Saleh, H. (2000). Internal structure analysis of asphalt mixes to improve the simulation of superpave gyratory compaction to field conditions.

Vavrik, W. R. (2000). *Asphalt mixture design concepts to develop aggregate interlock.* University of Illinois at Urbana-Champaign,

Vavrik, W. R., Pine, W. J., & Carpenter, S. H. (2002). Aggregate Blending for Asphalt Mix Design: Bailey Method. *Transportation Research Record, 1789*(1), 146-153. doi:10.3141/1789-16

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data, 3*(1), 9. doi:10.1186/s40537-016-0043-6

West, R., Rodezno, C., Leiva, F., & Yin, F. (2018). Development of a framework for balanced mix design. *Project NCHRP*, 20-07.

Xiao, Y. (2014). *Performance-based evaluation of unbound aggregates affecting mechanistic response and performance of flexible pavements.* University of Illinois at Urbana-Champaign,

Xiao, Y., & Tutumluer, E. (2017). Gradation and Packing Characteristics Affecting Stability of Granular Materials: Aggregate Imaging-Based Discrete Element Modeling Approach. *International Journal of Geomechanics, 17*(3). doi:10.1061/(asce)gm.1943-5622.0000735

Xiao, Y., Tutumluer, E., Qian, Y., & Siekmeier, J. (2012). Gradation Effects Influencing Mechanical Properties of Aggregate Base-Granular Subbase Materials in Minnesota. *Transportation Research Record: Journal of the Transportation Research Board, 2267*, 14-26. doi:10.3141/2267-02

Yeh, H.-h. (2007). *Random forests and gene selection to classify Arabidopsis thaliana ecotypes.* (M.S.). Michigan State University, Ann Arbor. ProQuest Dissertations & Theses Global database. (1448539)

Yin, F., Taylor, A. J., Tran, N., & Director, P. A. (2020). Performance Testing for Quality Control and Acceptance of Balanced Mix Design. In: NCAT.

Zavrtanik, N., Prosen, J., Tušar, M., & Turk, G. (2016). The use of artificial neural networks for modeling air void content in aggregate mixture. *Automation in Construction, 63*, 155-161. doi:10.1016/j.autcon.2015.12.009