This is the author's accepted manuscript (AAM) of :
Tong, B., Huang, W., Habbouche, J., Boz, I., Guo, Q., Diefenderfer, S. D., & Flintsch, G. W. (2025). Machine Learning–Based Prediction and Optimization of Balanced Mixture Design Performance Indices. Transportation Research Record, 0(0).
https://doi.org/10.1177/03611981251324209

Machine Learning–Based Prediction and Optimization of Balanced Mixture Design Performance Indices

Bilin Tong[1*], Wenjiang Huang[2*], Jhony Habbouche[3], Ilker Boz[4], Qing Guo[5], Stacey D. Diefenderfer[4], and Gerardo W. Flintsch[1]

[1]Virginia Tech Transportation Institute, Blacksburg, VA
[2]Department of Materials Science and Engineering, Johns Hopkins University, Baltimore, MD
[3]Asphalt Institute, Gilbert, AZ
[4]Virginia Transportation Research Council, Charlottesville, VA
[5]Department of Statistics, Virginia Tech, Blacksburg, VA
[*]These authors contributed equally to this work.

Corresponding Author: Bilin Tong, bilin@vt.edu

Machine Learning-Based Prediction and Optimization of Balanced Mixture Design Performance Indices

Bilin Tong*, Ph.D., Corresponding Author
Postdoctoral Associate
Virginia Tech Transportation Institute, Blacksburg, VA, 24060
Email: bilin@vt.edu, ORCID: 0000-0002-2831-8008

Wenjiang Huang*, Ph.D.
Postdoctoral Research Fellow
Material Science and Engineering, Johns Hopkins University, Baltimore, MD, 21218
Email: whuang75@jh.edu

Jhony Habbouche, Ph.D., P.E.
Western US Regional Engineer
Asphalt Institute, Gilbert, AZ, 85295
Email: jhabbouche@asphaltinstitute.org, ORCID: 0000-0002-6216-3134

Ilker Boz, Ph.D., P.E.
Research Scientist
Virginia Transportation Research Council, Charlottesville, VA, 22903
Email: Ilker.Boz@VDOT.Virginia.gov, ORCID: 0000-0002-8534-6996

Qing Guo, Ph.D.
Graduate Research Assistant
Department of Statistics, Virginia Tech, Blacksburg, VA, 24060
Email: qguo0701@vt.edu

Stacey D. Diefenderfer, Ph.D., P.E.
Associate Principal Research Scientist
Virginia Transportation Research Council, Charlottesville, VA, 22903
Email: stacey.diefenderfer@vdot.virginia.gov, ORCID: 0000-0002-6570-2121

Gerardo W. Flintsch, Ph.D., P.E.
Dan Pletta Professor
Virginia Tech Transportation Institute, Blacksburg, VA, 24060
Email: gflintsch@vtti.vt.edu, ORCID: 0000-0003-4440-0711

*These two authors contributed equally to this work.

ABSTRACT

The Balanced Mix Design (BMD) concept is an emerging methodology that facilitates the design of engineered asphalt mixtures. This approach is particularly beneficial for mixtures containing conventional and high reclaimed asphalt pavement contents, for which the traditional volumetric design methods may fail to effectively address the performance characteristics. However, due to production variability, these engineered mixtures can still fail to meet the required thresholds. Additionally, identifying the cause of this imbalance is challenging. To maximize the benefits of BMD implementation, this study introduces machine learning (ML) algorithms including linear regression (LR), random forest (RF), extreme gradient boosting (XGB), and support vector regression (SVR) as strategic tools to predict mixtures' BMD performance indices. 648 specimens fabricated for quality acceptance as part of the 2020 Virginia Accelerated Pavement Testing Program is used for the modeling and analysis. Durability, cracking, and rutting susceptibility of the specimens were evaluated using the Cantabro test, indirect tensile cracking test (IDT-CT), and Asphalt Pavement Analyzer (APA) rut test, respectively. Key outcomes include: (a) ML models, including RF, XGB, and SVR, demonstrated superior performance compared to LR; (b) Feature importance analysis from ML models identified dominant factors for each BMD test. Additionally, the reheating process was highlighted; (c) A pseudo in-situ deployment was simulated to optimize BMD implementation. The dimensionality reduction analysis — uniform manifold approximation and projection — highlighted the practical challenges associated with concurrently improving multiple performance metrics. Ultimately, the pivotal role of ML in advancing both the design and production phases was emphasized.

Keywords: machine learning, random forest, XGBoost, SVR, BMD, RAP, production variability.

1. INTRODUCTION

In the ever-evolving landscape of asphalt mixture design, the industry faces significant challenges driven by the necessity for high-performance pavements, the increasing demand for sustainable practices, and the escalating costs of raw materials (*1*). Current applications of the Superpave methodology have predominantly focused on volumetric properties of asphalt mixtures in the United State (U.S.). However, the dynamic nature of contemporary asphalt materials, characterized by the widespread use the extensive application of reclaimed asphalt pavement (RAP) and additives, necessitates a shift towards a design philosophy that emphasizes long-term performance.

The Balanced Mix Design (BMD) methodology has emerged as a pivotal innovation in this context. As delineated by the National Asphalt Pavement Association (NAPA) and the American Association of State Highway and Transportation Officials (AASHTO), BMD is characterized as an "*Asphalt mix design using performance tests on appropriately conditioned specimens that address multiple modes of distress taking into consideration, mix aging, traffic, climate, and location within the pavement structure* (*2, 3*)". The BMD methodology introduces performance testing criteria as a replacement for certain traditional aspects, such as volumetric properties, focusing on evaluating the resistance of mixtures to common distresses, such as rutting and cracking. To gain acceptance, BMD mixtures are required to meet defined performance benchmarks during design phase, with the possibility of additional performance tests during the production phase. Additionally, the integration of BMD methodology with the use of recycling agents (RAs) and/or softer binders, presents a strategic approach to design and production of engineered High RAP (HRAP) mixtures, providing an innovative and practical response to the demands of contemporary pavement construction (*4, 5*). In line with this advancement, a growing number of state highway agencies are actively investigating and implementing the BMD framework for the design and approval of asphalt mixtures (*5-7*).

Implementing the BMD framework typically involves two primary steps. The first step initiates with a design phase, wherein different state agencies may select varied foundational BMD tests and corresponding thresholds based on their specific requirements. For instance, in Virginia, three fast, simple, and practical performance-indicative tests were adopted for BMD surface mixtures (SMs). These tests include the Cantabro test, capped at a maximum Cantabro mass loss (CML) of 7.5%, to assess durability; the Indirect Tensile Cracking Test (IDT-CT) at intermediate temperatures, with a minimum cracking tolerance (CT) index of 70, for cracking performance assessment with short-term aged conditions; and the Asphalt Pavement Analyzer (APA) rut test, with a maximum rut depth (RD) limit of 8 mm, for rutting resistance evaluation (*8, 9*). The second step aims to verify that the design is reproduced during production by ensuring that the performance anticipated from the laboratory-designed mixtures is realized in the field through a quality assurance (QA) program. If performance criteria are not met (one or more indices exceed predetermined thresholds), the mixture is deemed unbalanced, thus diminishing the full potential benefits of BMD implementation (*10*).

Variations in asphalt mixture properties from the original design may arise during production and construction phases, because of the inherent variability in environmental conditions, plant operations, and natural materials (*10*). Additionally, the designed job mix formula

(JMF) is subjected to verification and potential plant modifications to account for actual production and on-site circumstances, irrespective of mixture design (*11*). This process poses significant challenges when applying the BMD methodology, which deviates from conventional volumetric property specification, thereby introducing potential difficulties for contractors. Furthermore, replicating the scale of mixture production that occurs in a plant setting, following laboratory design protocols, can be an intricate endeavor *(11)*. It is important to note that differences in asphalt mixture properties from the original design often trigger premature manifestations of pavement distress or can potentially result in outright structural failure *(10, 12)*. To identify and assess the sources, causes, and degrees of variability in gradation, volumetric, and mechanical properties, the National Cooperative Highway Research Program (NCHRP) Project 09-48 (Report 818) examined laboratory-mixed, laboratory-compacted specimens; plant-mixed, laboratory-compacted specimens used in volumetric acceptance testing; and plant-mixed, field-compacted specimens employed during density acceptance testing and forensic evaluation of in-place pavement. The aim of these assessments was to identify the sources of variability and to evaluate the precision and bias associated with the volumetric and mechanical properties of dense-graded asphalt mixtures *(12)*. Alternatively, the quality assessment can be done using reheated plant-mixed, laboratory-compacted specimens (reheated specimens), a practice observed in the Virginia Department of Transportation (VDOT) independent assurance practice (*13*).

In 2020, VTRC, VDOT, and Virginia Tech initiated collaborative research involving the design, placement, and laboratory and accelerated field testing of a control mixture and five BMD mixtures at the Virginia Accelerated Pavement Testing (APT) facility, owned by VDOT and situated at the Virginia Tech Transportation Institute (VTTI). This initiative aimed, in part, to apply VDOT BMD special provisions to production under highly controlled condition, offering an opportunity to scrutinize the implications of these specifications on different facets, including design, production, quality control and assurance practices, and construction of HRAP SMs *(4, 14)*.

In the current landscape, characterized by the exponential growth of data and the continuous advancement of iterative algorithms, the reemergence of machine learning (ML) approaches presents a promising avenue for enhancing the efficiency of analyzing data and identifying key factors. In the Virginia 2020 APT program, extensive fundamental performance tests on BMD samples in the field production were carried out, alongside analyses of the associated variability (*15*). Specimens fabricated for quality acceptance as part of the Virginia APT program were used for the ML modeling and analysis in this study.

2. OBJECTIVE AND SCOPE

The primary objective of this research was to develop ML models capable of predicting BMD performance indices based on mixture constituent and volumetric properties collected during both the design and field production phases at the project level. Four machine learning algorithms—linear regression (LR), random forest (RF), extreme gradient boosting (XGB), and support vector regression (SVR) —were employed to predict the performance of BMD mixtures. The performance of these different ML models was compared and analyzed. The secondary goal of this study was to use these developed ML models to assess the influence of constituent properties on the performance of BMD mixtures and to identify the dominant factors in BMD specimen

preparation. This was particularly challenging for traditional approaches due to the complex interactions of multiple constituent properties on the performance indices of asphalt mixtures. Finally, this study proposed the incorporation of ML into the design phase of BMD. To further demonstrate the application of ML models in BMD implementation, a simulated BMD deployment was strategically designed. Through simultaneous optimization across three primary BMD tests, optimized or ML-enhanced candidates were selected.

3. BACKGROUND AND LITERATURE REVIEW

3.1. Production Variability

The variability of different performance tests has been reported in many studies. Golalipour et al. (*1*) evaluated the performance of reheated specimens as part of the 2013 Federal Highway Administration's (FHWA's) Accelerated Loading Facility (ALF) study. This study determined coefficient of variance (COV) of six performance tests including: the Cantabro mass loss test, the N_{flex} factor test, the overlay test, the Illinois Flexibility Index Test (I-FIT) test, the IDT-CT test, and the cyclic fatigue test. The FI parameter from I-FIT presented the highest average COV (up to 34%) among all tests, while CT index achieved COV of 40% in Lane 7. Buttlar et al. (*16*) conducted cracking tests, including the Disk-Shaped Compact Tension Test (DCT), I-FIT test, and the IDEA-CT test, as part of the development of a local BMD method. Their findings indicated that while the average COV for DCT in field section evaluations was 19.5%, the I-FIT test showed a significantly higher average COV of 52.2%. In preparation for the full implementation of the BMD methodology, the VDOT initiated a round-robin study to assess testing repeatability across various laboratories. Forty-one laboratories, including those from the agencies, contractors, and independent testing facilities, participated in the initiative. Two distinct mix designs were utilized, with 46 sets of test specimens distributed for each design. Phase 1 of the study focused on evaluating IDEAL-CT test results from different labs, while Phase 2 commenced in 2021 (*17*). Additionally, currently accepted production tolerance of volumetric properties and gradation in Virginia can be found in (*18*). Current VDOT specified COV value of 18.3% for the CT index for 5 replicates.

Concerns about unbalanced performance in the resultant mixture, due to high variability in performance tests as reported in previous studies, have led researchers to dedicate efforts towards establishing tolerance limits for constituent properties during loose mixture sampling and specimen fabrication. However, recent research has found that a BMD mixture can still become unbalanced during production, even within specified tolerance limits. In the early stages of the BMD concept, Austerman et al. (19) modified asphalt binder content, gradation passing the No.200 sieve, and asphalt binder source in mixture to represent production variations. The study revealed that mixture performances, such as moisture susceptibility, may become unbalanced during production. They further noted that specific production factors can impact volumetric properties and the performance of mixtures developed using the BMD concept. Subsequently, Mogawer et al. (*7*) indicated that some mixtures produced at the lower limit of the optimal binder content (OBC) production tolerance, with or without alterations in aggregate gradation production tolerances, exhibited heightened susceptibility to fatigue cracking. Furthermore, it concluded that the binder source that can change during production results in unbalanced BMD mixtures. Finally, it was

recommended that a mixture's resistance to rutting be assessed at the upper production tolerance limit for OBC and for cracking at the following conditions: (1) the lower production tolerance limit for OBC; (2) the lower production tolerance limit for OBC in conjunction with the upper aggregate gradation production tolerance limits; and (3) the lower production tolerance limit for OBC combined with the lower aggregate gradation production tolerance limits. Recently, Bowers et al. (*20*) conducted an evaluation of fourteen accepted mixture designs, adjusting their coarse and fine gradation along with binder content to specified tolerance limits. This study found mixtures exhibited excellent rutting performance but displayed susceptibility to durability and cracking issues as the gradation and binder content change. It further underscored the potential for initially balanced mixtures to become unbalanced due to production variability. In addition to replicating production variability in the laboratory, Tong et al. (*15*) showcased field production variability for one Superpave control mixture and five BMD mixtures, designed using BMD Approach D (performance only), incorporating both conventional and high RAP content. Their investigation unveiled the possibility of encountering unbalanced BMD mixtures. It is important to note that in BMD Approach D, tolerance limits for volumetric properties are exempted except for binder content.

As a result, whether using BMD approaches with specified production tolerance or relying solely on performance-based BMD approach, concerns persist regarding the potential occurrence of unbalanced produced mixtures. Researchers may continue to propose refinements of BMD specifications, specifically, to determine an acceptable tolerance level (deviation from design) for gradation, asphalt content (AC), and volumetric properties, within which test values are still deemed to constitute a qualified or balanced mixture (*10*). However, simulating production variability in the laboratory and evaluating modified mixtures is a highly complex process, imposing significant costs in terms of manpower and materials. Additionally, efforts are also underway to model how component materials or mixtures properties influence corresponding mixture performance, as emphasized in several Virginia Transportation Research Council (VTRC) reports (*13, 20*). Yet, in a practical project, identifying the critical factors is highly intricate, as simple linear relationships may not adequately interpret the specific causes of production variability. This complexity arises due to the multitude of variables involved in production, underscoring the importance of employing nonlinear regression analyses.

3.2. Machine Learning Approaches

Some common ML approaches have been used in related prior studies (*21*). RF is a robust ensemble method that constructs multiple decision trees during the training phase. Each tree is developed using a randomly chosen subset of training data and features, enhancing model stability by mitigating overfitting and elevating prediction accuracy. Notably, RF effectively processes both categorical and continuous data and offers valuable insights into feature importance, aiding in the identification of critical predictors (*22*). XGB, another tree-based model, refines gradient boosting techniques by incorporating features such as built-in regularization and dynamic stopping criteria to counteract overfitting (*23)*. SVR, an extension of support vector machines (*24*), is designed to predict values within a specified deviation for all training data, with a focus on minimizing the upper bound of generalization errors, thereby enhancing model robustness. SVR utilizes kernel

functions to map input vectors into a higher-dimensional space, facilitating linear regression analysis and ensuring robust generalization performance across diverse datasets (*25*). The emergence of ML applications across a broad spectrum of industries exemplifies its wide-ranging utility. For example, in the context of asphalt concrete, Zeiada et al. (*26*) employed an artificial neural network (ANN) in conjunction with a forward sequential feature selection algorithm to identify the most influential design factors prevailing in warm climate regions using data sourced from the Long-Term Pavement Performance (LTPP) database. The efficacy of these applications can be attributed to the integration of feature importance, a main component of various ML techniques. Additionally, ML plays an increasingly vital role in predicting numerous essential parameters in asphalt concrete or asphalt pavement such as dynamic modulus (*27*), air void (*28*), rutting performance (*29*), fatigue performance (*30*), international roughness index (IRI) (*31*), and CT-index (*32*). Rahman et al. (*21*) explored the use of ML with large, diverse datasets to predict asphalt concrete metrics, such as IDT strength and Hamburg wheel tracking (HWT) test rut depth, marking a pivotal discussion on ML's integration into performance frameworks. They noted biases from trial batch samples and a significant data reduction due to missing values, decreasing datasets from 17,920 to 2,408 for HWT and 2,189 for IDT. Moreover, training ML models emphasized data quality over quantity, acknowledging that data variability often affects prediction accuracy, challenging the robustness and generalizability of models for BMD mixtures. Throughout the specimen preparation process, considerable variability can emerge from variations in technician handling, experimental conditions, and testing apparatus. This variability proves challenging to accurately incorporate as a quantifiable variable in the training of ML models. Therefore, conducting performance testing and data collection under rigorously controlled conditions may significantly mitigate the adverse effects on data quality resulting from disparate sources. Once a reliable ML model is developed, it can bridge the gap between the initial mix design and actual field performance by predicting the effects of mixture components and production variables on final product. Leveraging data fabricated from a highly controlled experiment schedule, ML algorithms can offer valuable insights into potential performance outcomes before field implementation, allowing for proactive adjustments and maximizing the benefits of BMD implementation.

4. METHODOLOGY

4.1. Experiments

A total of one control and five BMD 9.5mm dense-graded SMs were designed, produced, and sampled for evaluation in the 2020 Virginia APT program. The mixtures incorporated various combinations of three RAP contents, two binder performance grades (PGs), one RA, and one warm mix asphalt (WMA) additive. The RAP content in the mixtures was determined based on the total weight of the resulting mixtures. A 100% recycled binder availability was assumed, and the design asphalt contents accounted for the contribution of asphalt binder from RAP materials. The six mixtures are defined as follows:

- 30_C: a non-BMD mixture serving as a control with 30% RAP content and PG 64S-22 binder (with C denoting control and S denoting standard traffic). This mixture was

designed following VDOT's Superpave mix design methodology with the OBC selected for voids in total mixture (VTM) of 4%.

- 30_O: a BMD optimized version of mixture 30_C (with O denoting optimized) featuring the use of 30% RAP content and a PG 64S-22 at a relatively higher OBC when compared to mixture 30_C. The optimization process involved adjusting the aggregate gradation and increasing the asphalt binder content.
- 45_HR: a BMD optimized HRAP mixture with 45% RAP and PG 64S-22 binder. This mixture design resulted in a much higher OBC as compared with all other evaluated mixtures. No softer binder or RA were used during the design and production of this mixture.
- 45_HR_RA: a BMD optimized HRAP mixture with 45% RAP, PG 64S-22 binder, and an RA.
- 45_HR_L: a BMD optimized HRAP mixture with 45% RAP and a softer binder, PG 58-28.
- 60_HR_L_RA: a BMD HRAP mixture with 60% RAP, a softer binder (PG 58-28), and an RA.

The BMD optimized mixtures were designed following VDOT BMD special provision using Approach D (Performance Only) *(5, 6)*. The design volumetric properties and gradations of all mixtures are shown in Table 1. The BMD performance properties of the six mixtures at OBC determined during design are also shown in Table 1. More details about designed mixtures can be found in Tong et al. (*15*).

TABLE 1. Volumetric Properties and Gradations for APT Mixtures - Design

Mixture ID	**30_C**	**30_O**	**45_HR**	**45_HR_RA**	**45_HR_L**	**60_HR_L_RA**
Description	Non-BMD	BMD	BMD HRAP	BMD HRAP	BMD HRAP	BMD HRAP
Composition						
RAP Content, %	30	30	45	45	45	60
Asphalt Binder	PG 64S-22	PG 64S-22	PG 64S-22	PG 64S-22	PG 58-28	PG 58-28
Additives	WMA	WMA	WMA	WMA + RA	WMA	WMA + RA
Property						
N_{design}, gyrations	50	50	50	50	50	50
NMAS, mm	9.5	9.5	9.5	9.5	9.5	9.5
Asphalt Content, %	5.6	6.0	6.8	6.2	6.0	6.0
RBR	0.24	0.22	0.29	0.32	0.33	0.44
Rice SG (G_{mm})	2.531	2.517	2.497	2.519	2.521	2.538
VTM, %	4.0	2.1	5.2	2.7	4.2	1.5
VMA, %	16.8	15.9	20.2	16.9	17.9	15.3
VFA, %	76.2	87.0	74.3	84.0	76.6	90.0
FA Ratio	1.0	1.1	1.2	1.2	1.3	1.5
Mixture Bulk SG (G_{mb})	2.429	2.465	2.367	2.452	2.415	2.500
Performance Properties at Optimum Asphalt Binder Content (OBC)						
Cantabro Mass Loss, 25ºC, %	2.9	3.2	4.1	2.8	2.7	4.0
APA Rut Depth, 64ºC, mm	5.4	4.2	5.6	7.2	4.9	3.7
CT index, 25ºC	58	112	299	96	141	128
Gradation / Sieve Size	**% Passing**					
¾ in (19.0 mm)	100.0	100.0	100.0	100.0	100.0	100.0

½ in (12.5 mm)	100.0	100.0	100.0	100.0	100.0	100.0
3/8 in (9.5 mm)	94.0	96.0	93.0	93.0	92.0	92.0
No. 4 (4.75 mm)	66.0	66.0	63.0	63.0	63.0	62.0
No. 8 (2.36 mm)	39.0	41.0	38.0	38.0	40.0	39.0
No. 16 (1.18 mm)	--	--	--	--	--	--
No. 30 (600 µm)	23.0	17.0	18.0	18.0	19.0	20.0
No. 50 (300 µm)	--	--	--	--	--	--
No. 100 (150 µm)	--	--	--	--	--	--
No. 200 (75 µm)	5.8	6.6	7.4	7.4	7.5	8.3

S = standard traffic; BMD = balanced mix design; HRAP = high reclaimed asphalt pavement content mixture; RAP = reclaimed asphalt pavement; WMA = warm mix additive; RA = recycling agent; NMAS = nominal maximum aggregate size; RBR = recycled binder ratio, which is the ratio of RAP binder content to the mixture total binder content; SG = specific gravity; VTM = voids in total mixture; VMA = voids in mineral aggregate; VFA = voids filled with asphalt; FA = fines to aggregate; C = control; O = optimized; HR = high reclaimed asphalt pavement; RA = recycling agent; L = softer virgin binder.

4.2. BMD Testing

Three BMD tests adopted in Virginia (shown in Figure 1) are: Cantabro test, IDT-CT, and APA test, which assess durability, cracking, and rutting, respectively. The CML was determined in accordance with AASHTO T 401 (*33*) on three design specimens that were used to determine volumetric properties. The IDT-CT was conducted at a temperature of 25°C in accordance with ASTM D8225-19 (*34*) on five replicate test specimens compacted at 7±0.5% air voids. The APA test was performed in accordance with AASHTO T 340 (*33*) at a temperature of 64ºC; four replicate test specimens were tested in an APA Junior machine. The APA specimens were compacted at 7±0.5% air voids. VDOT BMD test thresholds were indicated in the **Introduction** section.

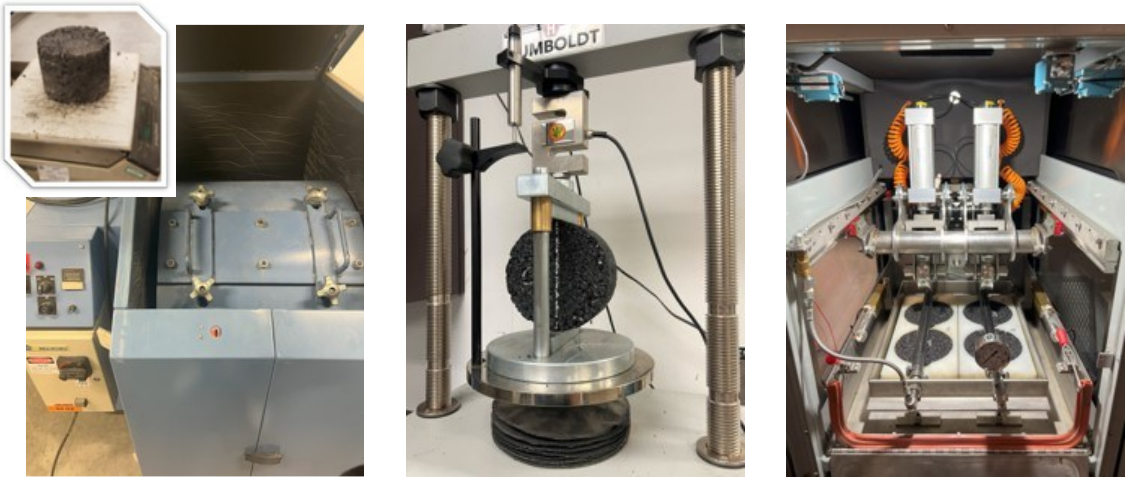

Figure 1. VDOT BMD Tests: (a) Cantabro Mass Loss Test; (b) IDT-CT; and (c) APA Rut Test. VDOT = Virginia Department of Transportation; BMD = balanced mix design; IDT-CT = indirect tensile cracking test; APA = asphalt pavement analyzer.

4.3. Production Sampling and Specimen Fabrication

A total of 648 specimens fabricated for quality acceptance as part of the 2020 Virginia APT experiment were used in this study. In addition to testing performance during the design process, a comprehensive laboratory assessment focusing on durability, cracking resistance, and rutting performance was conducted on mixtures collected during production. The test specimens included laboratory-mixed laboratory-compacted specimens (design), plant-mixed laboratory-compacted specimens (non-reheats), and plant-mixed laboratory-compacted specimens fabricated after a reheating process (reheats). Designs were laboratory-produced and laboratory-compacted, fabricated for mixture design purposes, while the produced mixtures were sampled to monitor their quality and associated production variability. Sampling of the loose mixture was conducted four times daily during production. The production sampling plan for each mixture is delineated in Figure 2. Samples were sequentially labeled from 1st to 4th (A through D) for each mixture. The loose mixture samples collected during production for BMD testing underwent one of two procedures: they were either transported to the producer's laboratory and promptly compacted into test specimens—these were referred to as plant specimens (non-reheats) - or they were placed in boxes, transported to VTRC, and stored in a climate-controlled area for the fabrication of reheated specimens at a later time (reheats) (*15*).

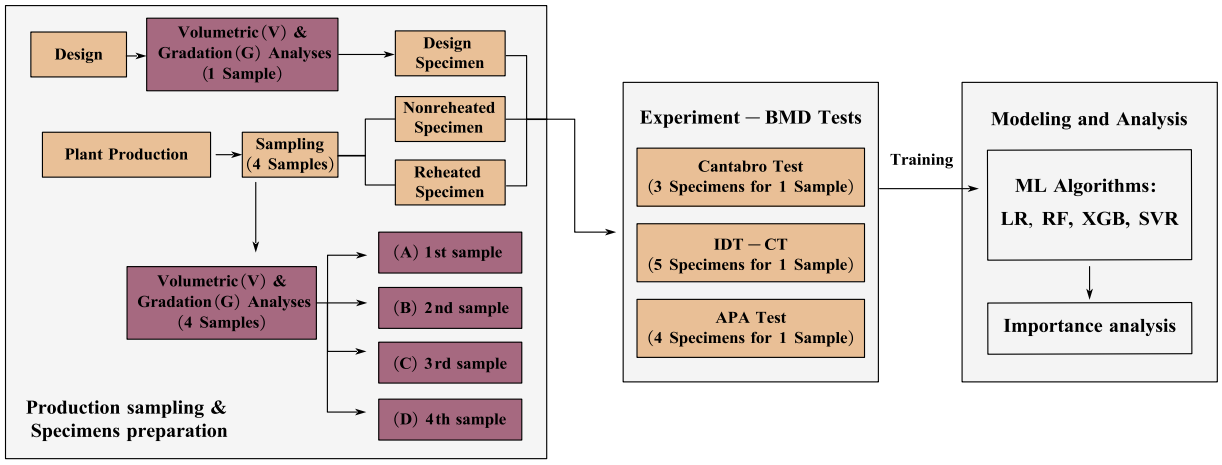

Figure 2. Experimental and Machine Learning (ML) Training Workflows. Left block: production sampling and specimen preparation. Middle block: balanced mix design (BMD) tests. Right block: ML models used in predicting three primary BMD tests. IDT-CT = indirect tensile cracking test; APA = asphalt pavement analyzer; G = aggregate gradation sieve analysis; V = volumetric properties analysis.

4.4. Modeling Approach

In this study, several ML algorithms were used to predict outcomes of three BMD tests in terms of CML, CT index, and APA rut depth. The employed models included LR (the simplest ML model), RF, XGB, and SVR. Figure 2 illustrates the experimental and ML training workflows. The left block displays the experimental procedures (detailed in the **Experiments** section), the middle block represents the three BMD tests (detailed in the **BMD testing** section), and the right block shows the ML algorithms used to predict those properties in BMD tests. The targets of ML prediction were the current VDOT adopted BMD mixture performance indices, which include CML, CT index, and APA rut depth. In developing the ML models, the dataset was randomly

divided into training and testing sets in an 80:20 ratio. Feature normalization was applied to the training set, with the same scaling coefficients subsequently used on the testing set. Hyperparameter optimization for each ML model was conducted via exhaustive grid-search method. For the RF model, hyperparameter variations included the number of trees (100 to 2000), tree depth (10 to 30), and minimum samples per leaf node (1 to 4). The XGB model parameters ranged from the number of trees (100 to 2000), learning rate (0.0001 to 0.1), tree depth (3 to 9), and subsample ratios of training instances and columns (0.7, 0.8, or 1). The SVR model's hyperparameters were the regularization parameter (0.001 to 100), epsilon (0.001 to 1), and kernel type (polynomial, radial basis, or sigmoid). The settings in an ML model that produced the highest mean R^2 values from the test set were selected as the optimal hyperparameters. The Scikit-learn library was utilized for constructing the ML models. Finally, to assess the influence of constituent properties on the performance of BMD mixtures and to identify the dominant factors in BMD specimen preparation, permutation importance, a model-agnostic method (*22, 35*), was utilized to interpret feature importance across all used ML models. This approach evaluates a feature's contribution to a model's predictive power by measuring performance changes when the feature's values are randomly shuffled, disrupting the correlation with the target variable. The comparative analysis of model performance, with and without the shuffled feature, quantifies each feature's impact.

The authors recognize that the data collected in this study was limited, but this method provided a highly controlled analysis environment and systematic data collection, reducing many avoidable issues such as data recording errors from different contractors or multiple laboratories. Predictions in a small dataset also demonstrate the effectiveness of ML and set an example for project-level analysis. Additionally, the approach and findings from this study can be adapted to accommodate other indices for mixture performance as different transportation agencies implement various BMD performance tests.

4.5. Training Features, Data Processing, and Evaluation Metrics

The input data collected for both designed and plant produced specimens contained 26 attributor/features, including categorical variables, volumetric properties, gradation parameters, and mechanical property. Categorical variables include design methods (encoded as 0 for conventional Superpave and 1 for BMD) and specimen types (encoded as 0 for design, 1 for non-reheats, and 2 for reheats). Volumetric properties consist of asphalt binder content, theoretical maximum and bulk specific gravities (G_{mm} and G_{mb}), VTM, voids in mineral aggregate (VMA), voids filled with asphalt (VFA), bulk and effective aggregate specific gravities (G_{sb} and G_{se}), fine aggregate to asphalt ratio (FA), percent binder absorbed (P_{ba}), effective binder content (P_{be}), effective asphalt film thickness (F_{be}), and RAP content (100% blending ratio is assumed). Gradation parameters are documented as the percentage passing through various sieve sizes: 12.5mm, 9.5mm, 4.75mm, 2.36mm, 1.18mm, 600µm, 300µm, 150µm, 75µm. The dynamic shear modulus (G*) at 64°C (high performance grade temperature), obtained from binder extracted and recovered from loose mixtures of field production sampling, was used as a mechanical property in this study. These features incorporated in the modeling for design and non-reheated specimens were recorded by VTRC, while those for non-reheated specimens were documented by the VDOT

district office. To address the challenges posed by limited data and remedy overfitting, Pearson correlation coefficients were calculated between each pair of features to determine the selection of eventual training features in ML models. This process, coupled with 5-fold cross-validation during training, ensures robust model performance, with further details provided in the **Results** section.

The datasets for each performance test comprised a different number of specimens: 162 for CML, 270 for IDT-CT, and 216 for APA RD tests. The APA tests were performed following AASHTO T-340 standard but only the average RD from two specimens on both left and right were reported, reducing the APA RD dataset to 108. Regarding missing data, the IDT-CT test had 3 missing entries, which were removed, while the APA test had two, which were filled using the median value of the corresponding RD. This approach to handling missing was empirically determined, considering the varying sizes of the dataset. Table 2 summarizes the details of each dataset.

TABLE 2. Data Sets/Specimens Used in Mahine Learning (ML) for Three Balanced Mix Design Tests

Data sets	CML	IDT-CT	APA RD test
Data sets/Specimens fabricated	162/162	270/270	108/216
Missing data sets	0	3	2
Pre-reserved data for pseudo in-site deployment	10	10	10
Data sets in an ML model	152	257	98

CML = Cantabro mass loss; IDT-CT = indirect tensile cracking test; APA = asphalt pavement analyzer; RD = rut depth.

To evaluate model performance, two metrics— the coefficient of determination (R^2) (Eq. 1) and the accuracy in terms of mean absolute percent error (*MAPE*) (Eq. 2) — were used,

$$R^2 = 1 - \left[\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}\right] \quad (1)$$

$$\text{Accuracy} = 1 - MAPE = 1 - \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i} \quad (2)$$

where y_i is the target values, $\bar{y}_i$ is the average of target values, $\hat{y}_i$ is the predicted value, and N is the number of data sets.

4.6. Simulating the BMD Deployment

Given the limited availability of materials for sampling during production and cost considerations, a pseudo in-situ deployment method is designed in this study to mimic real-world deployment scenarios for these ML models. For each dataset, 10 random specimens meeting specific criteria (CML < 7.5%, CT index > 70, APA < 8 mm) are pre-reserved and excluded from model development. Designing specimens in the laboratory that simultaneously satisfy these three primary criteria poses significant experimental challenges. This is because feature selection is highly empirical, and therefore the layered selection process based on various performance tests of interest often results in extensive repetitive work. Here, the aim of the ML pseudo in-situ deployment is to generate samples exhibiting all three balanced performances directly. ML models are trained with the remaining data (refer to Table 2) to predict each of these properties accurately (one ML model for one performance prediction). Once trained, samples are randomly generated

based on the mean and standard deviation of current dataset's distribution to ensure they can be created in-situ, and their BMD performances are predicted by the corresponding ML models. Each models' confidence is assessed using the corresponding 10 pre-reserved specimens. For instance, if an ML model predicts that 9 out of 10 pre-reserved CML samples have values less than 7.5%, the confidence level of this model is determined to be 90%. By establishing different thresholds for each predicted performance metric, it is possible to selectively identify samples that meet desired properties and confidence levels. Note that the confidence levels from the 10 pre-reserved specimens measure model precision against specific property thresholds, differing from the test set accuracy in training an ML model, which assesses overall predictive effectiveness across diverse data.

5. RESULTS AND DISCUSSION

5.1 Feature Selection

Pearson correlation coefficients (r) (ranging from -1 to 1 to indicate perfect negative and positive correlations respectively) were used for feature selection in this work. To mitigate multicollinearity, only features with an absolute r value less than 0.8 were chosen, resulting in 10 selected features. These include 2 categorical variables: "Design" and "Type", 4 constituent volumetric properties: AC, VTM, G_{se}, and the RAP content, 3 gradation parameters: sieve sizes of 75 μm, 600 μm and 2.36 mm, and 1 mechanical property: log |G*|. The log |G*| was measured at 64°C and 10 radians per second (rad/s) to account for binder viscosity suggested by Kim et al. (*35*) and NCHRP Report 871. The phase angle was excluded from the analysis for two key reasons. First, the phase angle δ exhibits a stronger linear relationship with G*, which could introduce collinearity in the feature selection process for ML models. Second, the objective was to select a single predictive parameter, rather than a combined one such as log|G*|/sinδ. Therefore, the feature selection rules applied in this work present a blend of both machine learning principles and domain knowledge. Figure 3, taking CT index as an instance, presents a correlation matrix for these features, with the lower triangle displaying linear regression plots and the upper triangle showing r values between features. The diagonal plots represent the distribution of each feature or the target property (the top left). Notably, a relatively high positive correlation (0.66) between the CT index and AC, and a negative correlation (-0.73) between the CT index and log|G*| were observed. These findings align with expectations that cracking resistance increases with higher AC content, while increased binder stiffness due to aging negatively impacts cracking resistance (*36*). While both AC and log|G*| show relatively higher correlations with CT index compared to other features, it is evident that they do not sufficiently capture the full complexity of CT index predictions. This highlights the necessity of employing machine learning algorithms to uncover the nonlinear interactions among the mixture properties and to enhance prediction accuracy.

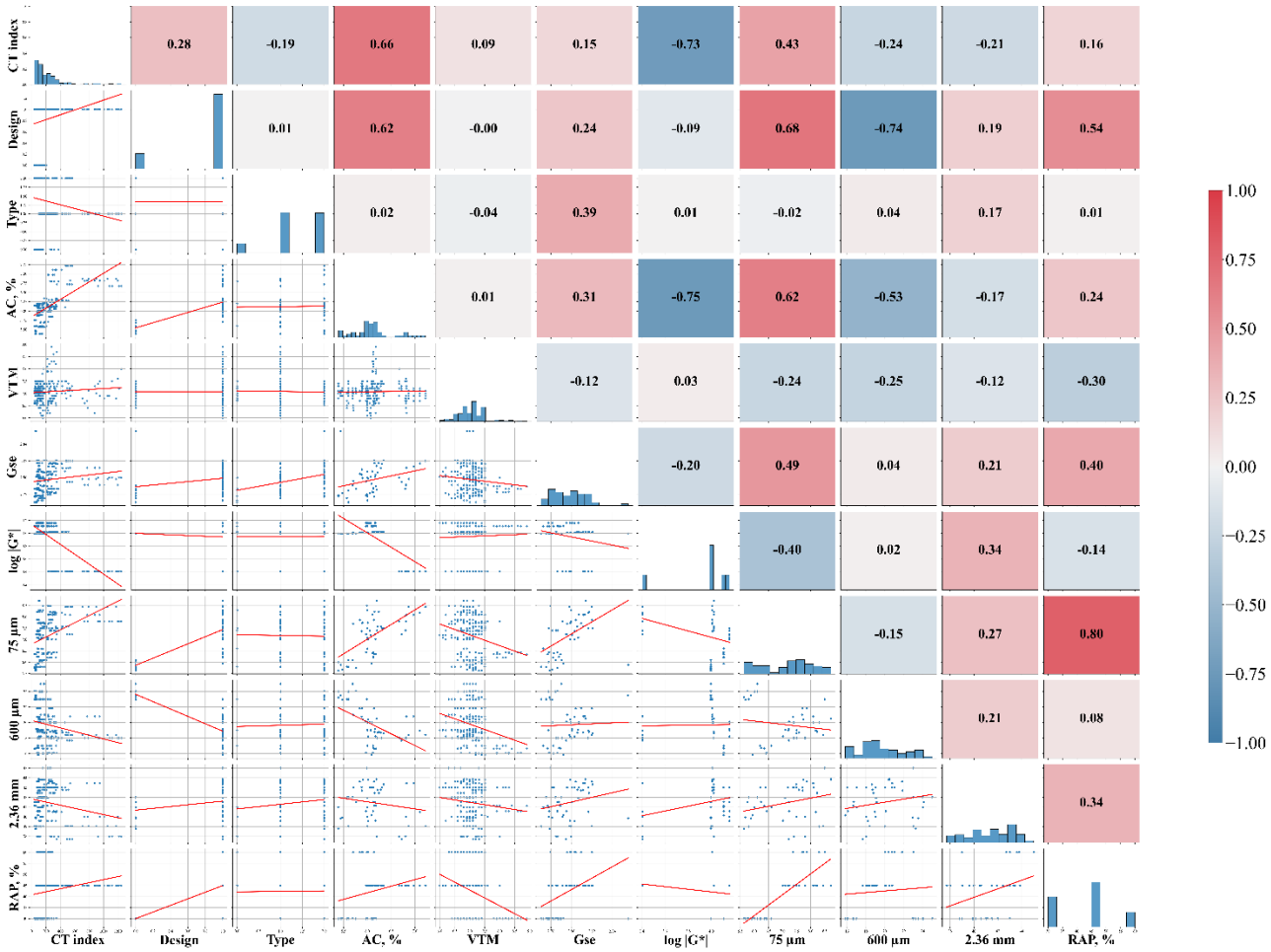

Figure 3. Correlation Matrix Derived from Pearson Correlation Coefficients (r) for Selected Features Associated with the IDT-CT Index. The lower triangle displays linear regression plots, while the upper triangle shows r values between features. The diagonal plots illustrate the distribution of each feature or the target property (e.g. top left).

5.2 Prediction Performance of ML Models

Figures 4, 5, and 6 display the predictive performance of the ML models for the CML, CT index, and APA RD parameters, respectively. After optimization of hyperparameters, the R^2 (Eq.1) and accuracy (Eq.2) values are presented for each model. Specifically, Figure 4 illustrates the performance for CML predictions with R^2 values of 0.86 (LR), 0.92 (RF), 0.94 (XGB), and 0.91 (SVR) and accuracy values of 0.88, 0.92, 0.92, and 0.90, respectively. Figure 5 details the CT index predictions with R^2 values of 0.63 (LR), 0.88 (RF), 0.92 (XGB), and 0.88 (SVR) and accuracy values of 0.47, 0.83, 0.81, and 0.77, respectively. Figure 6 shows the APA RD predictions with R^2 values of -0.03 (LR), 0.42 (RF), 0.56 (XGB), and 0.53 (SVR) and accuracy values of 0.69, 0.76, 0.79, and 0.77, respectively. The LR model consistently underperforms compared to the other ML models, which demonstrate stronger predictive abilities. Models like XGB consistently delivered outstanding R^2 and accuracy values for the CML and IDT-CT index, demonstrating their precision in predicting these two BMD performances. For the APA RD predictions, the low R^2 values indicate that the models do not closely follow the trends in the limited dataset (98 data

points). However, the relatively high accuracy scores suggest that the models effectively categorize most outcomes correctly.

Given the limited data available for each test, the division of data into training and testing sets can significantly affect the performance of the ML models. To evaluate the consistency of these models, a robustness check was conducted by randomly sampling the dataset using different seeds and training the models with the same optimized hyperparameters previously identified (i.e., from Figures 4, 5, and 6). This procedure was repeated 30 times to assess performance variability. Figure 7 illustrates the ML model capability (R^2 and accuracy) for BMD performance metrics prediction, and Table 3 summarizes the mean and standard deviation for each BMD performance metric. The LR model consistently underperforms relative to the other models, independently from data sampling. For CML and IDT-CT predictions, the models demonstrate stable prediction capability, as evidenced by the small standard deviation in R^2 (less than 0.2) and accuracy (less than 0.08). However, the APA RD predictions show significant fluctuations in R^2 values due to the limited dataset, though the accuracy fluctuation remains minimal, indicating that the models can reasonably categorize broader outcomes of the APA RD test.

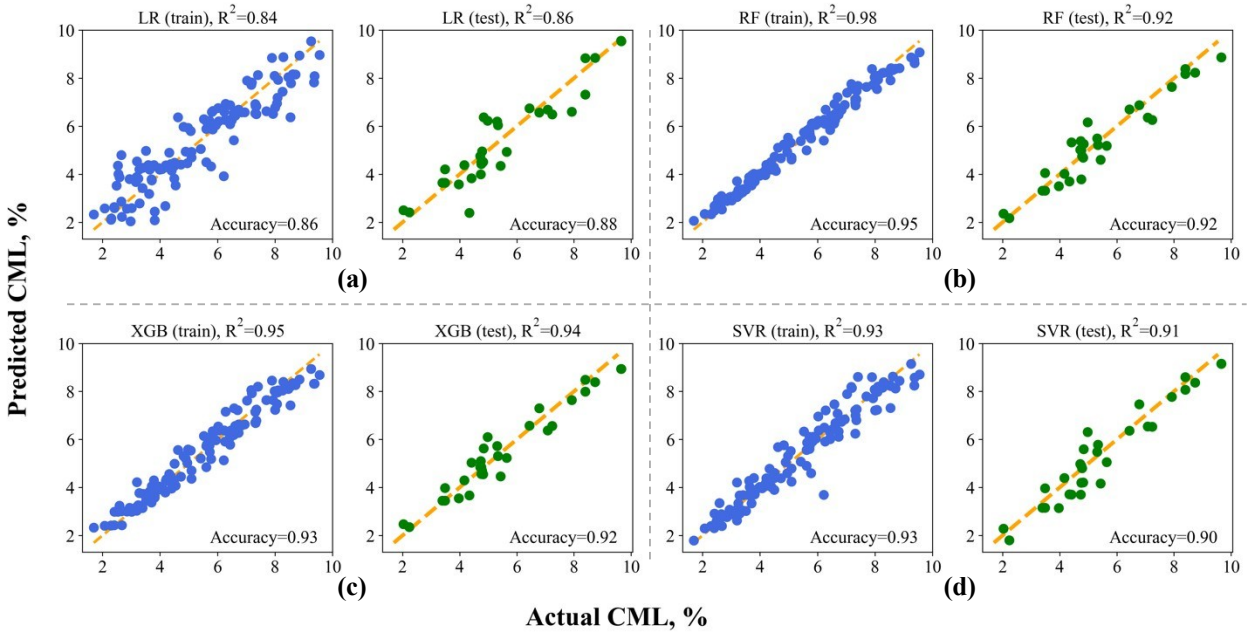

Figure 4. Prediction vs. Actual Cantabro mass loss (CML) using (a) LR, (b) RF, (c) XGB, and (d) SVR.

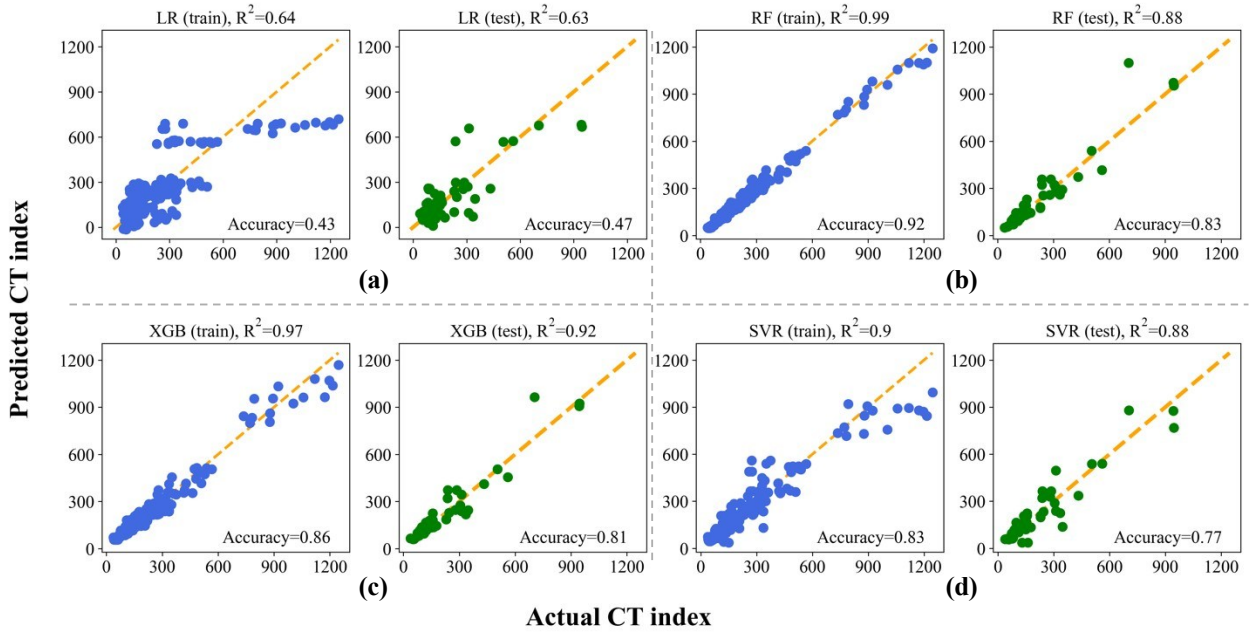

Figure 5. Prediction vs. Actual CT index using (a) LR, (b) RF, (c) XGB, and (d) SVR.

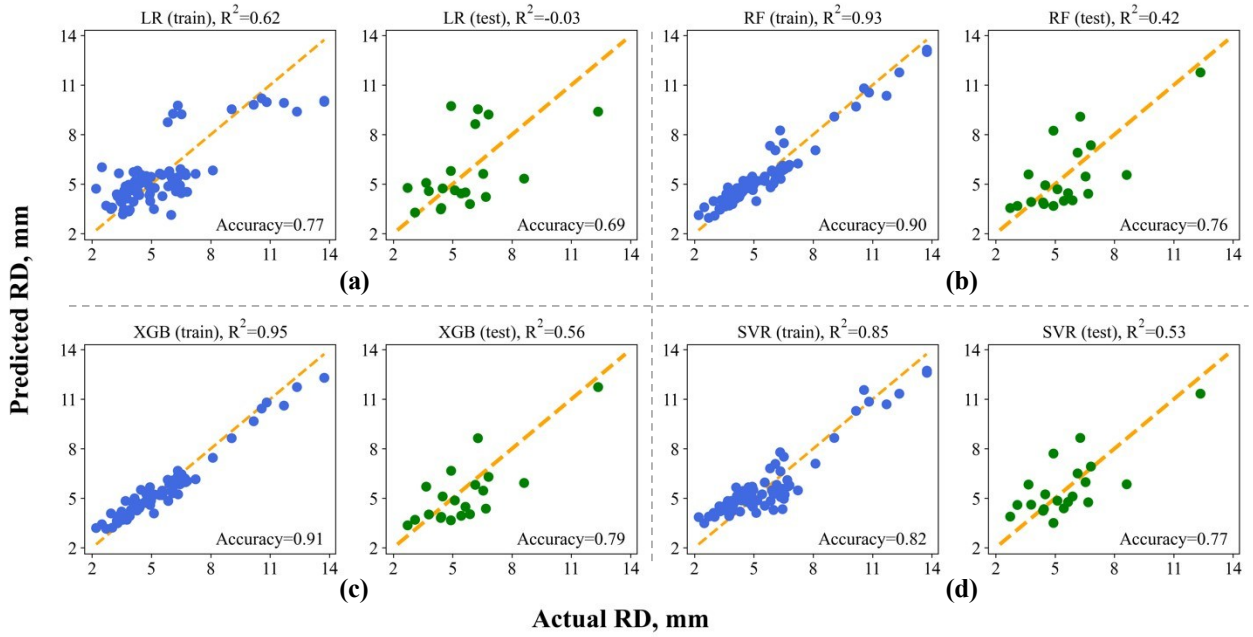

Figure 6. Prediction vs. Actual APA rut depth using (a) LR, (b) RF, (c) XGB, and (d) SVR.

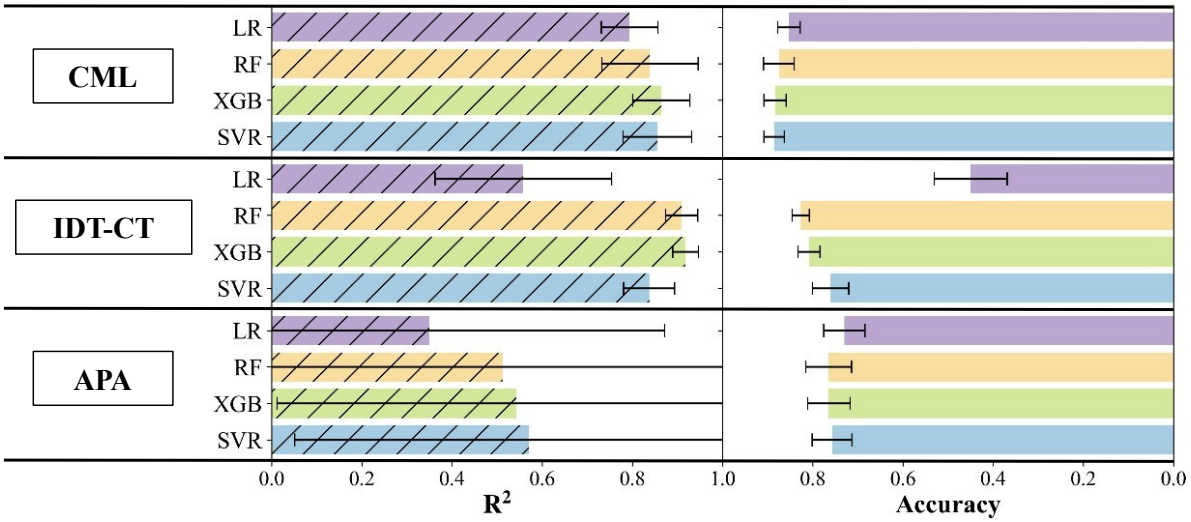

Figure 7. Variability in ML model performances (R^2 and accuracy) derived from 30 times random data splitting with different shuffle seeds.

TABLE 3. Mean and standard deviation in R^2 and accuracy from three BMD tests in Figure 7.

Model	Mean R^2 ± Std.			Mean accuracy ± Std.		
	CML	IDT	APA	CML	IDT	APA
LR	0.79 ± 0.06	0.56 ± 0.20	0.35 ± 0.52	0.85 ± 0.03	0.45 ± 0.08	0.73 ± 0.05
RF	0.84 ± 0.11	0.91 ± 0.04	0.51 ± 0.58	0.88 ± 0.03	0.83 ± 0.02	0.76 ± 0.05
XGB	0.86 ± 0.06	0.92 ± 0.03	0.54 ± 0.53	0.88 ± 0.02	0.81 ± 0.02	0.76 ± 0.05
SVR	0.86 ± 0.08	0.84 ± 0.06	0.57 ± 0.52	0.89 ± 0.02	0.76 ± 0.04	0.76 ± 0.04

CML = Cantabro mass loss; IDT-CT = indirect tensile cracking test; APA = asphalt pavement analyzer; RD = rut depth.

5.3 Feature Importance Analysis

Understanding the key factors influencing BMD tests allows engineers not only to focus on critical variables during the design phase but also to make proactive adjustments during the production phase. For this analysis, the models that demonstrated the highest accuracy across 30 instances of data sampling (Figure 7) were selected: XGB for CML (accuracy=0.93), RF for IDT-CT index (accuracy = 0.87), and RF for APA RD (accuracy = 0.83).

As shown in Figure 8, the feature importance analysis reveals distinct dominant factors influencing each BMD performance. For CML (Figure 8a), the key features include Type, 75 µm sieve size, AC and VTM. For IDT-CT index (Figure 8b), the primary features are AC, Type, 75 µm sieve size. For APA (Figure 8c), the dominant features are AC, log|G*| and VTM. Notably, the Type (type of specimens: design, non-reheats, and reheats) emerges as a significant feature for predicting both CML and IDT-CT index, suggesting the durability and cracking resistance may significantly vary by specimen type and be impacted by reheating effect. This observation is consistent with our previous studies emphasizing the impact of reheating on BMD performances (*15*). Furthermore, models on both the IDT-CT index and APA identified AC as a dominant factor, similarly highlighted in the previous study (*7*). The finding suggests that stricter control of asphalt binder content may be used as a proactive adjustment to reduce performance variability in mixture

performances during plant production. The log|G*| parameter, highlighted in (*35)* for reflecting binder viscosity—including contributions from virgin binder, RAP binder, and effects of aging—was also recognized by the ML models as a significant feature for the CT index and APA RD. This recognition further confirms the model's explanatory capability. Although it has been found that most gradation parameters do not significantly affect the performance of produced mixtures due to strict gradation controls at the plant (*15*), Austerman et al. (*19*) have suggested that greater attention should be paid to the percentage passing through the 75 µm sieve size during the production phase. Notably, 75 µm sieve size was identified as non-negligible in CML and CT index predictions, whereas the 600 µm and 2.36 mm sieve size showed limited impact on the model. These findings underscore the exceptional capability of the ML models to identify critical factors that align closely with prior observations and studies. Moreover, the VTM feature displays notable influences on both CML and APA RD predictions. While the impact of VTM deviation on CML tests has been reported in (*15*), its effect on APA RD tests may require more data for validation due to the unstable model prediction caused by limited APA data.

Design methods (i.e., Superpave or BMD) and RAP content are commonly emphasized. However, feature importance analysis reveals that neither the design method nor RAP content significantly impacts prediction performance. This is likely due to the use of BMD with a soft binder, with or without a recycling agent, which effectively compensates for variations in RAP content within the mixture (*4, 6*). Furthermore, the selection of other constituent properties in this study may effectively captures the differences introduced by BMD and Superpave design methods, clarifying the potential performance variations between HRAP mixtures and conventional RAP mixtures.

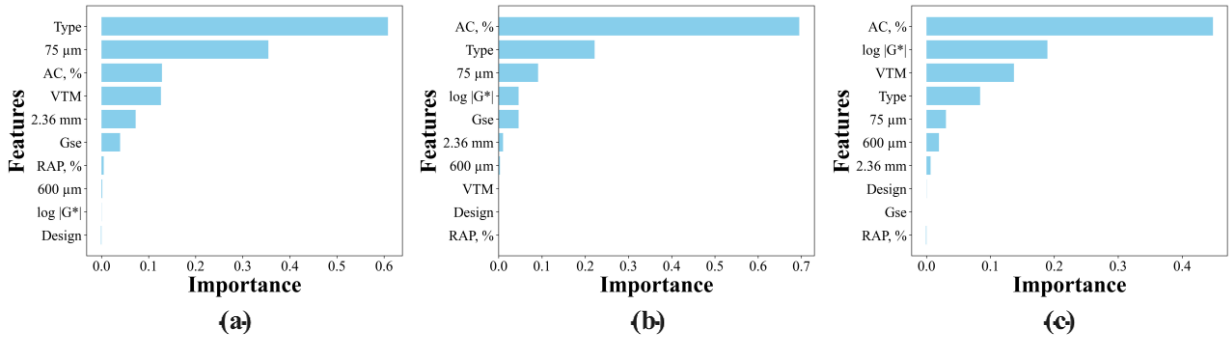

Figure 8. Feature importance from the permutation importance method for: (a) CML from XGB, (b) IDT-CT index from RF, and (c) APA RD from RF model.

5.4 Pseudo In-Situ Deployment

With the ML models demonstrating satisfactory predictive performance on BMD test properties, pseudo in-situ deployment was initiated to generate samples exhibiting balanced three primary BMD performances (i.e. CML < 7.5%, CT index > 70, APA < 8 mm). Figure 9 illustrates the schematic of this deployment process. Initially, 1000 synthetic samples were generated based on the mean and standard deviation of each feature from the original data pool, consisting of 618 data points excluding the 30 pre-reserved for testing. These samples were first subjected to filter-I, using the XGB model for CML prediction, to retain samples with CML values below 7.5%. Meanwhile, the CML values of 10 pre-reserved data were predicted to assess the confidence level

of filter-I. This approach was similarly applied in filter-II and filter-III for IDT-CT and APA models, with thresholds set at IDT-CT > 70 and APA RD < 8 mm, respectively. Samples passing all three filters were deemed balanced mixtures. In this evaluation, 812 samples passed filter-I, of which 811 proceeded past filter-II, and 771 passed filter-III. The confidence level for all filters reached 100%, indicating accurate predictions for all pre-reserved data.

It is commonly understood that simultaneously enhancing multiple primary BMD performances in mixture design is challenging. However, unlike earlier versions, contemporary asphalt materials are characterized by the widespread use of the extensive application of RAP and additives. This evolution complicates their performance and may challenge conventional wisdom (*4*). The utilization of ML models holds great potential to provide an enhanced solution for the three primary performance metrics. In this case, filter-IV applied stricter thresholds (CML < 5.8%, CT index > 200, and APA RD < 6 mm) to identify samples with enhanced mixture performances, narrowing the field to 60 candidates from the 771 remaining after filter-III. The stricter thresholds are based on the relationship between non-reheated and reheated specimens determined by previous study (*15*). Additionally, practical requirements for VTM content (7.0±0.5) in IDT and APA specimens led to the implementation of filter-V. After all five filters, only 19 out of the initial 1000 samples met the stringent criteria. This pseudo in-situ deployment demonstrates the effectiveness of using ML models to address the challenges of enhancing mixture properties, allowing for flexible adjustment of criteria in each filter to efficiently tailor the mixture design for specific practical applications.

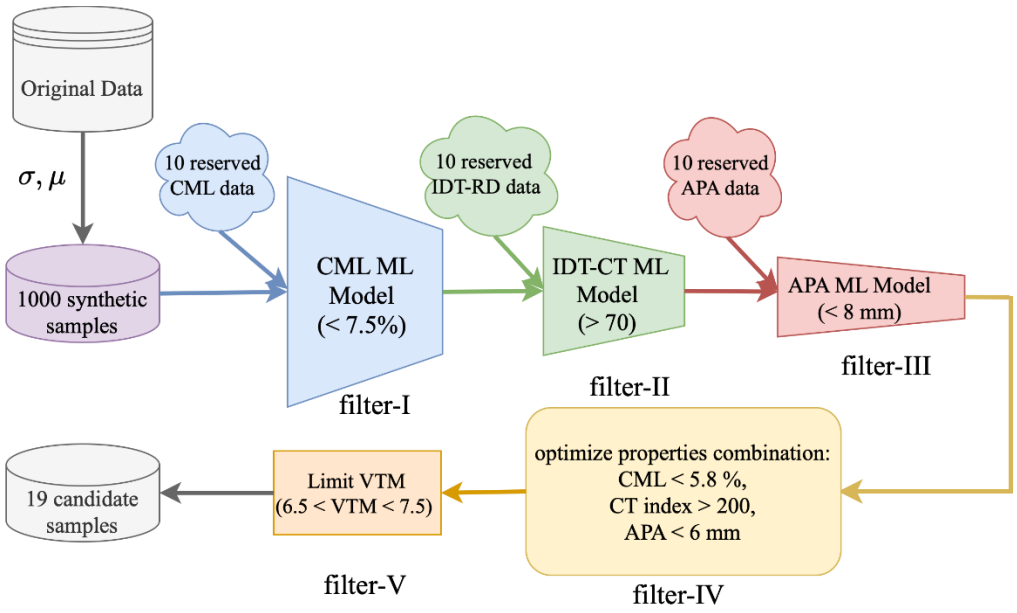

Figure 9. The schematic of pseudo in-situ deployment flowchart.

Experimentally designing a mixture that simultaneously enhances key properties—such as reducing CML and APA RD values while increasing the CT index—presents significant challenges due to the extensive feature space (10 features in this work). To understand this complexity, uniform manifold approximation and projection (UMAP) (*37*) was employed to

reduce the feature space into two dimensions for visualization and to compare the distribution of qualified samples from filter-IV (60 data points) against the original datasets. For each performance metric, such as CML, data were divided into two groups: those with CML values below 5.8 and those with higher values. This categorization was similarly applied to the CT and APA datasets. Note that the samples from filter-IV, which have all three primary performances enhanced, were plotted alongside the original datasets, where each dataset represents only one BMD performance. Figure 10 displays these distributions, with qualified samples marked by purple stars and the original dataset's qualified data points represented by orange dots in Figures 10a, 10b, and 10c for CML, CT index, and APA RD test, respectively. The visualization revealed that samples meeting all enhanced criteria rarely overlapped with the distributions of the original datasets. This highlights the challenges engineers face in achieving all desired enhancements experimentally, as the target distribution with all enhanced performances is scarce within the broad feature space, making exhaustive searches in experiments typically impractical.

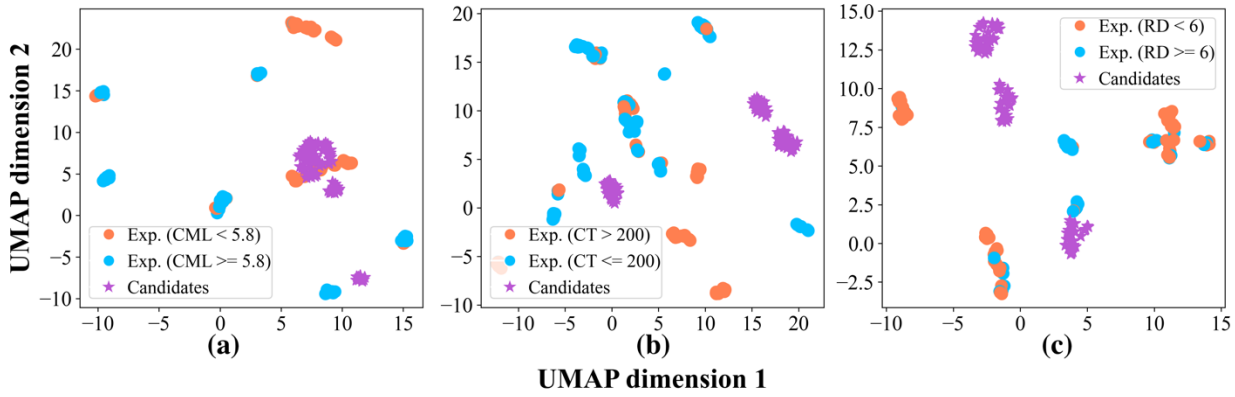

Figure 10. Uniform Manifold Approximation and Projection (UMAP) reduces the feature space to two dimensions for (a) CML, (b) CT index, and (c) APA RD datasets. The number of neighbors used in UMAP is 10, and it is asserted that the choice of hyperparameters has a negligible impact on the conclusions drawn.

5.5. Discussion

This section summarizes the potential applications of ML models within the BMD framework during the design and production phases, thereby showcasing the enhancements brought by ML models. Figure 11 illustrates the current BMD framework, using BMD approach D as an example. A pivotal loop in the BMD design phase involves the selection of feature combinations based on required performance tests. Each loop is not only subject to variability from specimen fabrication, testing, and device, but may also become cumbersome due to the need to restart the selection process if the results of a performance test fail to meet the specified BMD thresholds. With the assistance of developed ML models, this cumbersome design loop can potentially be bypassed, significantly enhancing design efficiency. Additionally, concerns about unbalanced cases resulting from mixture production can be addressed through project-level ML models, which identify dominant factors contributing to production variance. It is suggested that more stringent control of the constituent properties could reduce performance variability in produced mixtures. Furthermore, if ML models can be trained with more data in a highly controlled environment and further

validation of models' predictions can be achieved, it is highly promising that the QA program in BMD implementation will soon significantly reduce the reliance on sampling mixtures for BMD performance tests during production, opting instead to predict performance based solely on component properties. Finally, this study establishes a framework for the application of ML in BMD and verifies its feasibility at the project level. State DOT agencies can reference the data collection methods presented in this study to build a comprehensive database, facilitating the development of reliable ML models that incorporate a wider variety of material sources, such as binder types, aggregate resources, and additives.

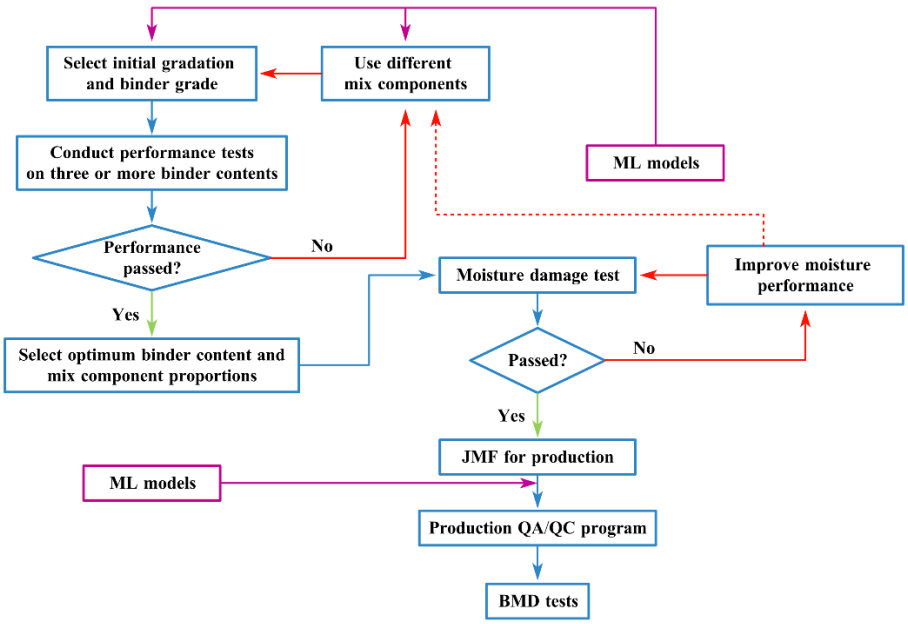

Figure 11. Explore the Use of ML Models to Support the Implementation of BMD.

6. CONCLUSION

This study leverages ML algorithms to help address some of the challenges prevalent in the existing BMD framework, thereby enhancing its implementation during both the design and production phases. Four prominent ML algorithms were employed to predict three primary performance metrics of BMD mixtures, namely, CML, CT index and APA RD. Hyperparameters for each ML model were optimized using an exhaustive grid search method with sufficient parameter spaces. Among the models, RF, XGB, and SVR exhibited superior performance compared to LR, effectively capturing the complex interactions within the mixture properties. The robustness of the ML models was evaluated and validated through 30 independent training/testing dataset splits in a random manner. ML models reliably delivered accurate predictions across all BMD tests, corroborated by the identification of significant influencing factors such as Type of specimens, AC, log|G*|, 75 μm sieve size passing percentage, and VTM. The impact of the reheating process on mixtures' durability and cracking performance was also identified, supported by previous studies. The study also recommended stricter control of identified features to minimize performance variability in mixtures. However, the models faced challenges in achieving stable R^2

values in the APA RD test, likely due to limited data availability and the use of average values from APA specimens.

To demonstrate the practical application of the developed ML models in BMD mixture design, a pseudo in-situ deployment (Figure 9) was strategically implemented to filter and select 60 out of 1,000 synthetic samples with optimal BMD performances through simultaneous optimization of three BMD performance metrics. By simply adjusting the filters in the pseudo in-situ deployment, one can effortlessly fine-tune the desired BMD performances. Dimensionality reduction via the UMAP method visualized minimal overlap in the distributions of these 60 candidates compared to the original datasets, underscoring the complexities of managing large feature spaces and improving multiple BMD performances simultaneously in experimental settings. Current study establishes a framework for implementing ML in BMD applications and demonstrates its feasibility at the project level. State DOT agencies can utilize the data collection methodologies outlined in this research to construct a comprehensive database. This database will support the development of robust ML models, incorporating a broader range of material sources such as binder types, aggregate resources, and additives.

ACKNOWLEDGMENTS

The authors thank Andrew Barbour, Troy Deeds, Donnie Dodds, Derek Lister, and Jennifer Samuels of VTRC; Danny Martinez Rodriguez formerly of VTRC; Billy Hobbs and Kenny Smith of VTTI; and Fabrizio Meroni and Max Ratcliffe, formerly of VTTI, for their outstanding efforts in sample collection and testing. Appreciation is also extended to Travis Higgs and Clyde Landreth of VDOT Salem District for their assistance and support. Sincere appreciation is expressed to Boxley Asphalt Paving and Ingevity for their contributions to this effort.

FUNDING

The author(s) received the following financial support for the research, authorship, and/or publication of this article: Virginia Transportation Research Council and Virginia Department of Transportation sponsored this study (Grant UPC 124200, https://vtrc.virginia.gov/projects/all-projects/124200/.

DISCLAIMER

The contents of this paper reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the Virginia Department of Transportation, the Commonwealth Transportation Board, or the Federal Highway Administration. This paper does not constitute a standard, specification, or regulation. Any inclusion of manufacturer names, trade names, or trademarks is for identification purposes only and is not to be considered an endorsement.

AUTHOR CONTRIBUTION STATEMENT

The authors confirm contribution to the paper as follows: conceptualization, formal analysis, methodology, interpretation of result, validation, and draft manuscript preparation: B. Tong, W. Huang, J. Habbouche, I. Boz, Q. Guo, S.D. Diefenderfer, and G.W. Flintsch. Modeling and visualization: B. Tong, W. Huang. Project administration, supervision: J. Habbouche, I. Boz, S.D.

Diefenderfer, and G.W. Flintsch. All authors reviewed the results and approved the final version of the manuscript. B. Tong and W. Huang contributed equally to this work.

CONFLICTS OF INTEREST

Jhony Habbouche is a member of Transportation Research Record's Editorial Board. All other authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

DATA AVAILABILITY STATEMENT

The data presented in this manuscript can be accessed upon request from and approval of the corresponding author and funding agency. The data is also available in the final report of project VTRC/VDOT 124200 approved and currently under edits by the funding agency.

REFERENCES

1. Golalipour, A., V. Veginati, and D.J. Mensching. Evaluation of Asphalt Mixture Performance Using Cracking and Durability Tests at a Full-Scale Pavement Facility. *Transportation Research Record*, 2021. 2675: 226 - 236.
2. National Center for Asphalt Technology and National Asphalt Pavement Association. Balanced Mix Design Resource Guide. https://www.asphaltpavement.org/expertise/engineering/resources/bmd-resource-guide. Accessed January 24, 2023.
3. American Association of State Highway and Transportation Officials (*AASHTO*). Standard Specifications for Transportation Materials and Methods of Sampling and Testing. AASHTO, Washington, D.C., 2021.
4. Tong, B., J. Habbouche, S.D. Diefenderfer, and G.W. Flintsch. Multi-level performance evaluation of BMD surface mixtures with conventional and high RAP contents: a case study in Virginia. *International Journal of Pavement Engineering*, 2024.
5. West, R., C. Rodezno, F. Leiva, and F. Yin. *NCHRP Project 20-07/Task 406 Development of a Framework for Balanced Mix Design*. *Transportation Research Board*, Washington, DC, 2018.
6. Yin, F., and R. West. *Balanced Mix Design Resource Guide*. NAPA IS-143, National Asphalt Pavement Association, Greenbelt, MD, 2021.
7. Mogawer, W.S., A.J. Austerman, K.D. Stuart, I.A. Abdalfattah, and F. Zhou. Comprehensive Balanced Mixture Design Protocol: Design Phase through Production Quality Assurance. *Transportation Research Record*. 2023.
8. Bowers, B. F., Diefenderfer, S. D., Moore, N. and Lynn, T. (2022) Balanced Mix Design and Benchmarking: A Case Study in Establishing Performance Test Thresholds. *Transportation Research Record*, 2676 (12): 586–598.
9. Habbouche, J., I. Boz, and S. D. Diefenderfer. Validation of Performance-Based Specifications for Surface Asphalt Mixtures in Virginia. *Transportation Research Record*, 2021. 2676 (5): 277 – 296.

10. Lynn Wright, T.Y., B.F. Bowers, S.D. Diefenderfer, and I. Boz. Evaluation of the Influence of Production Tolerance Limits on Virginia Balanced Mixtures. *Transportation Research Record*. 2024.
11. Brown, E. R., P. S. Kandhal, F.L. Roberts, Y. R. Kim, D Lee, and T.W. Kennedy. Hot Mix Asphalt, Materials, Mixture Design and Construction, 3rd Edition, *NAPA Research and Education Foundation*. Lanham, Maryland, 2009
12. Mohammad, L.N., M.A. Elseifi, S.B. Cooper, C.S. Hughes, J.W. Button, and E. Dukatz Comparing the Volumetric and Mechanical Properties of Laboratory and Field Specimens of Asphalt Concrete. *NCHRP Report*. 2016.
13. Diefenderfer, S.D., J. Habbouche, and I. Boz. *Balanced Mix Design for Surface Asphalt Mixtures: 2020 Field Trials*. Publication *VTRC 23-R13*. Virginia Transportation Research Council, Charlottesville, 2023.
14. Tong, B., J. Habbouche, B. K. Diefenderfer, and G. W. Flintsch. Rutting Performance Evaluation of BMD Surface Mixtures with Conventional and High RAP Contents under Full-Scale Accelerated Testing. *Materials*, 16, 7611. https://doi.org/10.3390/ma16247611, 2023.
15. Tong, B., J. Habbouche, S.D. Diefenderfer, and G.W. Flintsch. Assessment of Production Variability and Composite Performance Index for Conventional and High RAP BMD Mixtures, *Transportation Research Record*, 2024.
16. Buttlar, W. G., L. Urra-Contreras, B. Jahangiri, P. Rath, and H. Majidifard. Support for balanced asphalt mixture design specification development in Missouri. *CMR 20-010*. Springfield, MO: Missouri DOT, 2020
17. Habbouche, J., I. Boz, and S. D. Diefenderfer. Interlaboratory Study for the Indirect Tensile Cracking Test at Intermediate Temperature: Phase II. *VTRC Report 23-R3*. Virginia Transportation Research Council, Charlottesville, 2022.
18. VDOT. Road and Bridge Specification. Virginia Department of Transportation, 2020.
19. Austerman, A. J., W. S. Mogawer, and K. D. Stuart. Influence of Production Considerations on Balanced Mixture Designs. *Transportation Research Record*, 2672(28), 426-437. https://doi.org/10.1177/0361198118786826, 2018.
20. Bowers, B. F., T. Lynn., F. Yin., N. Moore., S.D. Diefenderfer., I. Boz. Impact of Production Variability on Balanced Mix Designs in Virginia. *FHWA/VTRC 23-R20*. 2023. URL : https://rosap.ntl.bts.gov/view/dot/67963
21. Rahman, S., A. Bhasin, and A.D. Smit. Exploring the use of machine learning to predict metrics related to asphalt mixture performance. *Construction and Building Materials*, 295, 123585, 2021
22. Breiman, Leo. Random forests. *Machine learning*, 45(1):5–32, 2001.
23. Chen, T., and C. Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794), 2016
24. Vapnik, V., S. Golowich, and A. Smola. Support vector method for function approximation, regression estimation and signal processing. *Advances in neural information processing systems*, 9, 1996.

25. Smola, A. J., and B. Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14, 199-222, 2004.
26. Zeiada, W., S. A. Dabous, K. Hamad, R. Al-Ruzouq, and M. A. Khalil. Machine learning for pavement performance modelling in warm climate regions. *Arabian journal for science and engineering*, 45(5), 4091-4109, 2020.
27. Daneshvar, D., and A. Behnood. Estimation of the dynamic modulus of asphalt concretes using random forests algorithm. *International Journal of Pavement Engineering*, 23(2), 250-260, 2022.
28. Zavrtanik, N., J. Prosen, M. Tušar, and G. Turk. The use of artificial neural networks for modeling air void content in aggregate mixture. *Automation in Construction*, 63, 155-161. 2016.
29. Liu, J., F. Liu, C. Zheng, D. Zhou, and L. Wang. Optimizing asphalt mix design through predicting the rut depth of asphalt pavement using machine learning. *Construction and Building Materials*, 356, 129211, 2022.
30. Ahmed, T. M., P. L. Green, and H. A. Khalid. Predicting fatigue performance of hot mix asphalt using artificial neural networks. *Road Materials and Pavement Design*, 18(sup2), 141-154, 2017
31. Zhou, Q., E. Okte, and I. L. Al-Qadi. Predicting pavement roughness using deep learning algorithms. *Transportation Research Record*, 2675(11), 1062-1072, 2021.
32. Nguyen, L. N., T. H. Le, L. Q. Nguyen, and V. Q. Tran. Machine learning approaches for predicting Cracking Tolerance Index (CT Index) of asphalt concrete containing reclaimed asphalt pavement. *Plos one*, 18(10), e0287255, 2023.
33. American Association of State Highway and Transportation Officials. *Standard Specifications for Transportation Materials and Methods of Sampling and Testing, and AASHTO Provisional Standards.* Washington, DC, 2017.
34. ASTM International. *ASTM D 8225-19: Standard Test Method for Determination of Cracking Tolerance Index of Asphalt Mixture Using the Indirect Tensile Cracking Test at Intermediate Temperature*. West Conshohocken, PA, 2019.
35. Kim, Y.R., C. Castorena, M.D. Elwardany, F.Y. Rad, S. Underwood, A. Gundha, P. Gudipudi, M.J. Farrar, and R.R. Glaser. Long-Term Aging of Asphalt Mixtures for Performance Testing and Prediction: Phase III Results, 2021.
36. Fisher, A., C. Rudin, and F. Dominici. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of machine learning research : JMLR*, 20. 2018.
37. McInnes, L., J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv*:1802.03426. 2018.