CIVIL ENGINEERING STUDIES
Illinois Center for Transportation Series No. 23-027
UILU-ENG-2023-2027
ISSN: 0197-9191

Review of Illinois Multiple Stress Creep and Recovery Data for Future Implementation

Prepared By
Ramez Hajj
Babak Asadi
University of Illinois Urbana-Champaign

Research Report No. FHWA-ICT-23-020

A report of the findings of

ICT PROJECT R27-SP63
Data Analysis and Review of Multiple Stress Creep and Recovery vs. Performance-graded High-Temperature Test Results, Including Binder Elastic Behavior/Response

https://doi.org/10.36501/0197-9191/23-027

Illinois Center for Transportation

December 2023

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. FHWA-ICT-23-020	**2. Government Accession No.** N/A	**3. Recipient's Catalog No.** N/A
4. Title and Subtitle Review of Illinois Multiple Stress Creep and Recovery Data for Future Implementation		**5. Report Date** December 2023 **6. Performing Organization Code** N/A
7. Authors Ramez Hajj, https://orcid.org/0000-0003-0579-5618 Babak Asadi, https://orcid.org/0000-0002-4946-8793		**8. Performing Organization Report No.** ICT-23-027 UILU-2023-2027
9. Performing Organization Name and Address Illinois Center for Transportation Department of Civil and Environmental Engineering University of Illinois at Urbana-Champaign 205 North Mathews Avenue, MC-250 Urbana, IL 61801		**10. Work Unit No.** N/A **11. Contract or Grant No.** R27-SP63
12. Sponsoring Agency Name and Address Illinois Department of Transportation (SPR) Bureau of Research 126 East Ash Street Springfield, IL 62704		**13. Type of Report and Period Covered** Final Report 4/1/23–12/31/23 **14. Sponsoring Agency Code**

15. Supplementary Notes
Conducted in cooperation with the U.S. Department of Transportation, Federal Highway Administration.
https://doi.org/10.36501/0197-9191/23-027

16. Abstract
The multiple stress creep recovery (MSCR) test, per AASHTO M 332, or as a supplement to M 320, has been widely implemented across the United States and in other parts of the world for assessing the rutting resistance of asphalt binder. Meanwhile, the Illinois Department of Transportation (IDOT) has conducted wide-ranging MSCR testing since 2006 but has not yet implemented this test into the binder specification, which utilizes AASHTO M 320 with other PG Plus tests for modified asphalt binder. This report presents a review of the existing literature and an update of specifications in the United States for the rutting resistance of asphalt binder. The report also presents an analysis of existing MSCR data collected by IDOT and a machine learning model with the ability to predict asphalt binder elastic recovery using MSCR. This model demonstrated the potential effectiveness of MSCR to serve as an appropriate surrogate for this test, but further study is needed of Illinois mixture performance compared to MSCR to best understand how this specification can be implemented for local materials and traffic levels.

17. Key Words Asphalt, Asphalt Binder, MSCR, Rutting, Elastic Recovery, Machine Learning	**18. Distribution Statement** No restrictions. This document is available through the National Technical Information Service, Springfield, VA 22161.		
19. Security Classif. (of this report) Unclassified	**20. Security Classif. (of this page)** Unclassified	**21. No. of Pages** 48	**22. Price** N/A

Form DOT F 1700.7 (8-72) **Reproduction of completed page authorized**

ACKNOWLEDGMENT, DISCLAIMER, MANUFACTURERS' NAMES

This publication is based on the results of **ICT-R27-SP63: Data Analysis and Review of Multiple Stress Creep and Recovery vs. Performance-graded High-temperature Test Results, Including Binder Elastic Behavior/Response.** ICT-R27-SP63 was conducted in cooperation with the Illinois Center for Transportation; the Illinois Department of Transportation; and the U.S. Department of Transportation, Federal Highway Administration.

Members of the Technical Review Panel (TRP) were the following:

- Kelly Morse, TRP Chair, Illinois Department of Transportation
- Dennis Bachman, Federal Highway Administration
- Kevin Burke, Illinois Asphalt Pavement Association
- Cody Davidson, Illinois Department of Transportation
- Justin Grant, Illinois Department of Transportation
- Brian Hill, Illinois Department of Transportation
- Brian Pfeifer, Illinois Department of Transportation
- Ronald Price, Illinois Department of Transportation
- William C. Snyder, Illinois Department of Transportation
- Megan Swanson, Illinois Department of Transportation
- Jason Wielinski, Asphalt Institute

EXECUTIVE SUMMARY

Rutting, or permanent deformation, is one of the most critical pavement distresses in flexible pavements. The causes of rutting can be many, but one of the most critical issues in terms of rutting prevention is the proper selection of asphalt component materials, including the asphalt binder. Over the last two decades, the conventional rutting parameter used by road agencies, G*/sinδ, measured at the high Performance Grade (PG) temperature at 10 rad/s, is insufficient for properly characterizing asphalt binders. To this end, many agencies have implemented "PG Plus" tests, such as elastic recovery using a ductilometer. However, these tests are burdensome and may not be the most effective for quantifying high-temperature resistance of the asphalt binder because they are run closer to the binder's intermediate temperature.

To this end, many agencies have adopted the AASHTO M 332 specification, which includes the standard PG grading as well as the multiple stress creep recovery (MSCR) test. MSCR applies repeated creep-recovery loading at two stress levels, 0.1 kPa, and 3.2 kPa, with 1 second of creep loading and 9 seconds of recovery time per cycle, using the existing dynamic shear rheometer (DSR) device. Some agencies run this test at 64℃ for all types of asphalt binder, while others run it at a high PG temperature. The M 332 specification assigns each asphalt binder an appropriate traffic level and speed for use in addition to the climatic grading assigned by conventional PG. In addition, some agencies continue to use AASHTO M 320, the standard specification for performance grading of asphalt binder, but have added MSCR as a PG Plus test.

The present study conducted a review of the existing literature on MSCR implementation in the United States and around the world. The literature review revealed that a majority of states have implemented MSCR in some form into their binder specification, but that there are remaining optimizations to be made with respect to the test. Findings are overall mixed in terms of which parameters are most effective between the non-recoverable creep compliance (J_{nr}), the percent recovery observed (%R), and the stress sensitivity between the two stress levels (J_{nrdiff}). Particularly, J_{nrdiff} generally had little relationship with mixture test results, despite its wide adoption. The review also revealed that using additional stress levels, extending the recovery time, and other modifications to the test could be more useful in capturing actual recovery ability of a binder at high temperatures.

The second stage of research consisted of a data review of Illinois Department of Transportation's existing MSCR dataset. IDOT has collected extensive MSCR data since 2006 across a wide range of asphalt binder sources, mostly among polymer-modified asphalt binder. After removal of outliers, %R at both stress levels had the best relationship directly to the elastic recovery test results, and both J_{nr} and %R had almost perfect relationships between the two stress levels. However, the relationship between *%R* and elastic recovery was not sufficient for developing a specification based on this parameter alone, so the research team employed bagging and boosting machine learning techniques to predict elastic recovery from all MSCR inputs plus the high PG. The extra trees and extreme gradient boosting (XGBoost) methods predicted the elastic recovery very well compared to other existing tests, with %R at both stress levels serving as the most important factor using both methods, and high PG and J_{nrdiff} having less impact on binder elastic recovery.

Overall, this project demonstrated the promise of MSCR, but further study is needed before implementation for Illinois-specific materials. The research team suggests conducting a comprehensive mixture study using the Hamburg wheel-tracking device to validate MSCR test methods, thresholds, and parameters using IDOT mixtures. They also suggest conducting a more comprehensive study of unmodified binder and binder modified with softeners in alignment with IDOT's new binder specification, which allows for these types of modifiers to be used when blended at the terminal.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW

NEED FOR SUPPLEMENTAL TEST TO *G**/SINδ CRITERION

Rutting, a prominent distress in flexible pavements, presents as longitudinal depressions within the wheel path. Severe rutting not only compromises driver comfort, but also reduces the longevity of asphalt pavements, resulting in heightened safety concerns for road users. This distress phenomenon primarily arises from the accumulation of permanent deformation within asphalt layers, primarily induced by repetitive vehicular loads, notably those imposed by heavy, slow-moving traffic. Extensive research has established a direct correlation between pavement distresses and the rheological properties of asphalt binder. Consequently, there is a burgeoning interest in developing a performance-based rutting indicator that can be seamlessly integrated into the asphalt binder grading system.

In the past, binder susceptibility to permanent deformation was typically assessed using traditional test methods like the ring and ball softening point. During the Strategic Highway Research Plan studies, Anderson and Kennedy (1993) proposed that $G^*/\sin\delta$ as measured in the dynamic shear rheometer (DSR) could be a better indicator of how well binder resists rutting at various temperatures, a conclusion primarily based on the performance of unmodified binder. Measuring this parameter within the linear range, employing just a few cycles of sinusoidal loading, does not properly capture the asphalt binder's impact on rutting within the mixture, particularly in the case of modified binder (D'Angelo et al., 2007). This has led to initiatives aimed at developing new testing methods capable of predicting performance irrespective of the type of modification (Bahia et al. 2001).

To ensure the enhancement of polymer modification in high-quality asphalt binder, state and local highway agencies have required conducting one or more additional Performance Grade (PG) Plus tests (e.g., elastic recovery (ER) or toughness and tenacity) alongside the existing PG asphalt binder specification. The development of the multiple stress creep recovery (MSCR) test arose from the necessity for a standardized procedure to assess both the elastic response and the existence of polymer modifiers in asphalt binder as a surrogate to the ER test (D'Angelo et al., 2007). The use of MSCR over the conventional PG Plus tests offers notable advantages. First, although it shares similarities with the ER test in its ability to detect the presence of elastomeric modifiers, the MSCR test provides more precise insights into the behavior and effectiveness of the polymer network within the asphalt binder. Moreover, the MSCR test requires no additional equipment beyond the DSR device, which is already utilized in many laboratories to perform AASHTO T 315 asphalt binder testing. Additionally, it involves less preparation time and significantly reduces testing time compared to the ER test, which typically takes several hours to complete.

HISTORY OF MSCR DEVELOPMENT

The MSCR test protocol underwent distinct phases in its development. Inspired by the repeated creep and recovery test proposed by Bahia et al. (2001), the MSCR test applies 11 stress levels ranging from 25 to 25,600 Pa to evaluate how asphalt binders' permanent deformation response differs at various

stress levels (D'Angelo et al., 2007). This test introduced a novel parameter called non-recoverable creep compliance (J_{nr}), which demonstrated a significantly stronger correlation with rutting observed on Accelerated Loading Facility sections compared to the Superpave rutting parameter, $G^*/\sin\delta$. Later, the MSCR test was adopted as a specification test for asphalt binder aged in a rolling thin-film oven (RTFO). This version involved two stress levels, 0.1 kPa and 3.2 kPa, with 10 repetitions of the creep-recovery loading pattern (1-second creep followed by 9-second recovery) at each stress level. Percent recovery (%*R*) was also introduced to assess the elastic response and stress dependence of the tested asphalt binder. In this phase, a relationship between J_{nr} and the Superpave binder criteria was established, leading to the development of an improved grading method. The final phase included modifications to enhance the accuracy and reliability of the test, such as increasing the number of creep and recovery cycles at 0.1 kPa from 10 to 20 to reach a steady-state condition. Additionally, the standard grade criterion for J_{nr} was adjusted to 4.5 kPa^{-1} (AASHTO M 332). Table 1 presents a catalog of the four classifications outlined in AASHTO M 332–14 and ASTM D8239–18. With the implementation of MSCR, the conventional practice of grade bumping can be abandoned due to its incapacity to accurately mirror real performance characteristics under specific pavement design temperature conditions. Note that revising this table is a primary objective of any future specification.

Table 1. MSCR Grading System

Grade	Maximum Non-Recoverable Creep Compliance at 3.2 kPa	Traffic Condition
S	4.5	Traffic Level < 10 million ESALs* and Traffic Speed > 70 km/h (43.5 mph)
H	2	Traffic Level 10–30 million ESALs or Traffic Speed 20–70 km/h (12.4 – 43.5 mph)
V	1	Traffic Level > 30 million ESALs or Traffic Speed < 20 km/h (12.4 mph)
E	0.5	Traffic Level > 30 million ESALs and Traffic Speed < 20 km/h (12.4 mph)

* ESAL stands for equivalent single axle load.

CALCULATION OF J_{NR} AND %*R*

The two primary parameters derived from the MSCR test are the non-recoverable creep compliance (J_{nr}) and the percentage of recovery (%*R*). In order to gain a more comprehensive grasp of J_{nr} and %*R*, a schematic representation of the strain response of an asphalt binder is provided in Figure 1 as an illustrative demonstration. To maintain conciseness, Figure 1 exclusively displays strain information from an arbitrary cycle, recorded either at 0.1 kPa or 3.2 kPa.

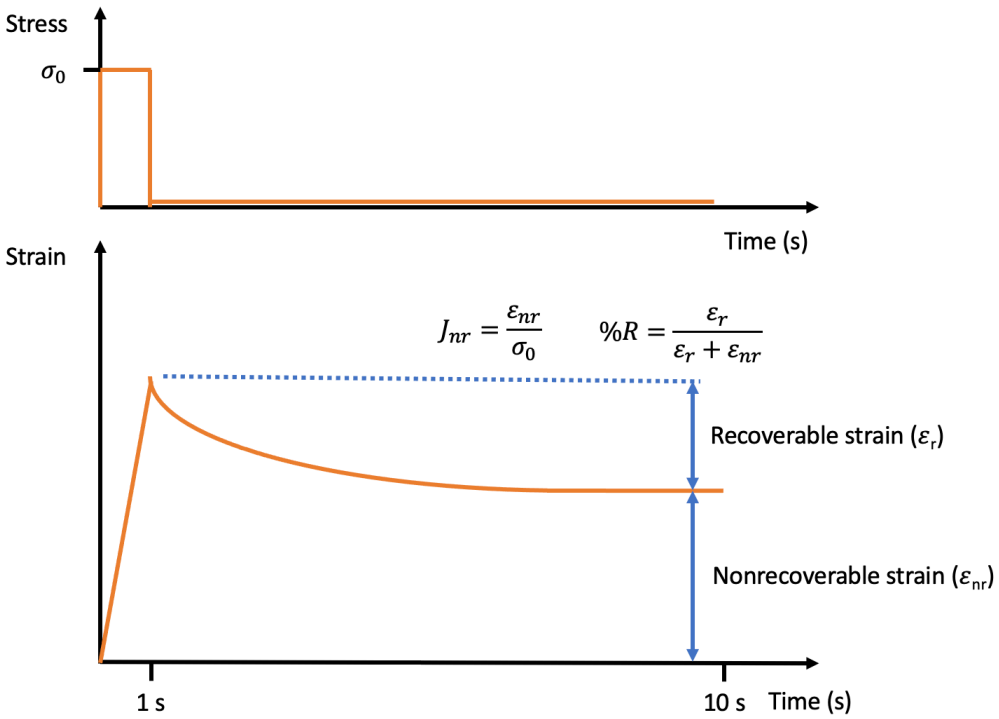

Figure 1. Graph. Sample of single cycle of MSCR test and associated parameters.

Within this context, ε_0 represents the initial shear strain at the onset of a creep phase, ε_c corresponds to the shear strain at the conclusion of the creep phase, and ε_r signifies the recoverable shear strain at the termination of the recovery phase. The calculation of non-recoverable creep compliance involves dividing the strain that remains unrecovered at the conclusion of the recovery phase by the applied shear stress, as indicated in the equation in Figure 2. On the other hand, the percentage recovery is ascertained by taking the ratio of the strain that has been recuperated at the end of the recovery phase to the shear strain induced at the end of the creep phase, as depicted in the equation in Figure 3.

$$J_{nr}(\sigma, N) = \frac{\varepsilon_{nr}}{\sigma}$$

Figure 2. Equation. Non-recoverable creep compliance equation.

$$R(\sigma, N) = \frac{\varepsilon_r}{\varepsilon_r + \varepsilon_{nr}}$$

Figure 3. Equation. Percentage recovery equation.

in which σ denotes the applied stress measured in kPa (i.e., 0.1 and 3.2 kPa) and N represents the count of creep and recovery cycles of interest. Following the application of equations presented in Figure 2 and Figure 3 for each cycle, the average non-recoverable creep compliance and average percentage recovery are computed for each stress level. This computation is performed to

accommodate the variations observed in the measured values. As per the MSCR test protocol, 20 creep and recovery cycles are carried out at a stress level of 0.1 kPa, while 10 creep and recovery cycles are conducted at a stress level of 3.2 kPa. The initial 10 cycles at the lower stress level of 0.1 kPa serve to condition the specimen, with the subsequent 10 cycles dedicated for data analysis. The average non-recoverable creep compliance ($J_{nr0.1}$) and the average percent recovery ($R_{0.1}$) at a stress level of 0.1 kPa are calculated using equations in Figures 4 and 5, respectively. Likewise, for a stress level of 3.2 kPa, $J_{nr3.2}$ and $R_{3.2}$ are determined with equations in Figures 6 and 7, respectively.

$$J_{nr0.1} = \frac{\sum_{N=11}^{20} J_{nr}(0.1, N)}{10}$$

Figure 4. Equation. Average non-recoverable creep compliance at 0.1 kPa.

$$R_{0.1} = \frac{\sum_{N=11}^{20} R(0.1, N)}{10}$$

Figure 5. Equation. Average percent recovery at 0.1 kPa.

$$J_{nr3.2} = \frac{\sum_{N=1}^{10} J_{nr}(3.2, N)}{10}$$

Figure 6. Equation. Average non-recoverable creep compliance at 3.2 kPa.

$$R_{3.2} = \frac{\sum_{N=1}^{10} R(3.2, N)}{10}$$

Figure 7. Equation. Average percent recovery at 3.2 kPa.

CORRELATION OF MSCR WITH MIXTURE PERFORMANCE

Ever since the introduction of the MSCR test, significant efforts have been dedicated to investigating its correlation with asphalt mixture rutting performance through various field and laboratory studies. While field verification studies solely relied on rut depth (D'Angelo et al., 2007; D'Angelo, 2009), laboratory-based research has employed a diverse array of rutting indicators, dependent on the specific mixture rutting test chosen. These included the Hamburg wheel-tracking device test (Wasage et al., 2011; Zhang et al., 2015; Hajj et al., 2019), the unconfined dynamic creep test (Tabatabaee & Tabatabaee, 2010), the flow number test (de Barros et al., 2022; Behnood et al., 2016), and the stress sweep rutting test (de Barros et al., 2022), among others. Table 2 summarizes many of these studies in detail to provide a sample of the available literature.

It is important to note that some studies found little to no relationship between MSCR parameters and mixture performance. For example, Salim et al. (2019) observed no significant relationship between %*R* at 3.2 kPa and observed a superior performance of both $J_{nr3.2}$ and the Superpave rutting parameter. The same research also observed a relationship between %*R* at 3.2 kPa with fatigue resistance (Kaloush et al., 2019), which is not usually a used purpose but does relate to the conventional ER test.

Table 2. A Sample of Studies Presenting Correlation between MSCR Test and Rutting Performance of Asphalt Concrete

Asphalt Binder	Stress Level	Temp (°C)	Rutting Test	Temp (°C)/ Air Void (%)	MSCR/ AC Rutting Test Parameters	Correlation	Reference
Crosslinked Styrene Butadiene polymer modifier all with PG-HT 58	11 levels from 25 to 26,500 Pa	58.0-70.1	Hamburg Wheel-Tracking Test (HWTT)	58.0-70.0 (7.0% air void)	J_{nr} at 12,800 Pa/ 10,000 cycles rut	J_{nr} = 0.442x - 0.521 (R^2=0.93) G*/sin(δ) = -0.084x+3.235 (R^2=0.65)	D'Angelo (2009)
Polymer content 2-5% (SBS, EVA, and commercial)	11 levels from 25 to 26,500 Pa	60.0	French Wheel Tracking Test	60.0 (7.2% air void)	J_{nr} at 26,500 Pa/ French rutting @ 30,000 cycles	J_{nr} = 0.094x – 0.26 (R^2=0.77) G*/sin = -30.55ln(x) + 59.29 (R^2=0.27)	Dreessen et al. (2009)
PG 58-22 modified with 3%-15% crumb rubber (CRM)	100 and 3,200 Pa	52.0-88.0	Unconfined Dynamic Creep	40 ± 1 (4 ± 1% air void)	J_{nr} at 100 and 3,200 Pa/ Unconfined Dynamic Creep	R^2 =0.86 for J_{nr} at 100 Pa R^2=0.83 for J_{nr} at 3200 pa R^2=0.74 for G*/sin	Tabatabaee and Tabatabaee (2010)
SBS and CRM-modified binder	11 levels from 25 to 26,500 Pa	40.0-60.0	HWTT	40.0-60.0 (~4% air void)	J_{nr} at 12,800 Pa/ rut depth after 10,000 cycles	ln(J_{nr}) = 0.0011 ln(x) (R^2=0.98)	Wasage et al. (2011)
PG70-XX or higher	100 and 3,200 Pa	64.0-76.0	HWTT and Repeated Loading Permanent Deformation (RLPD)	50.0 for HWTT and 40.0 for RLPD (7 ± 1% air void)	J_{nr} at 0.1 and 3.2 KPa/ HWTT rut depth at 10,000 passes; J_{nr} at 0.1 and 3.2 KPa/ RLPD accumulated at 10,000 cycles	HWTT @ 0.1 KPa: R^2 of 0.83, 0.85, and 0.77 at 64, 70, and 76°C and roughly the same values @ 3.2 KPa R^2=0.44 between HWTT and G*/sin RLPD: @ 0.1 KPa: R^2 of 0.92, 0.94, and 0.96 at 64, 70, and 76°C and the corresponding values of 0.88, 0.81, and 0.72 @ 3.2 KPa R^2=0.59 between HWTT and G*/sin	Zhang et al. (2015)

Asphalt Binder	Stress Level	Temp (°C)	Rutting Test	Temp (°C)/ Air Void (%)	MSCR/ AC Rutting Test Parameters	Correlation	Reference
SBS and ground tire rubber-modified binder	100 and 3,200 Pa	64.0	Flow Number (FN) Test	51.0 (7 ± 0.5% air void)	J_{nr} at 0.1 and 3.2 KPa/ flow number	R^2 =0.57 for J_{nr} at 100 Pa R^2=0.71 for J_{nr} at 3200 pa R^2=0.48 and 0.37 for G*/sin at Unaged and RTFO-aged, respectively	Behnood et al. (2016)
SBS, CRM, and EVA-modified binder	9 levels from 100 to 26,500 Pa	60.0	Wheel Tracking Test	60.0 (7 ± 0.5% air void)	J_{nr} at 0.1 and 0.8, 3.2 and 25.6 KPa/ rut depth after 20,000 cycles (40,000 passes)	Rut depth = $2.49x^{0.3855}$ for J_{nr} at 3200 Pa (R^2=0.99) Rut depth = $22.51x^{-0.284}$ for G*/sin (R^2=0.43)	Radhakrishnan et al. (2018)
PG 64-22 from three different sources	100 and 3,200 Pa	58.0 and 64.0	HWTT	50.0 (7 ± 1% air void)	$R_{0.1}$, $R_{3.2}$, R_{diff}, $J_{nr0.1}$, $J_{nr3.2}$, and J_{nrdif} / HWTT parameters including rut depth	$R_{0.1}$: R^2 =0.49 and 0.33 (power) at 58 and 64°C; $R_{3.2}$: R^2 =0.52 and 0.38 (power) at 58 and 64°C; R_{diff}: R^2 =0.58 and 0.54 (Exponential) at 58 and 64°C; $J_{nr0.1}$: R^2 =0.57 and 0.30 (Exponential) at 58 and 64°C; $J_{nr3.2}$: R^2 =0.55 and 0.32 (Exponential) at 58 and 64°C; J_{nrdiff}: R^2 =0.55 (Linear) and 0.99 (Power) at 58 and 64°C	Walubita et al. (2022)
Highly polymer-modified	5 levels from 100 to 26,500 Pa	82.0	HWTT and FN	50.0 (7 ± 1% air void)	J_{nr} at 6.4, 12.5, and 25.6 kPa/ HWTT rut depth J_{nr} at 6.4, 12.5, and 25.6 kPa/ flow number	HWTT: @ 6.4: J_{nr} = 0.74x-1.46 (R^2=0.99) @ 12.5: J_{nr} = 1.66x-3.47 (R^2=0.93) @ 25.6: J_{nr} = 2.43x-4.76 (R^2=0.87) FN: @ 6.4: J_{nr} = -0.0003x +1.18 (R^2=0.72) @ 12.5: J_{nr} = -0.0006x + 2.17 (R^2=0.57) @ 25.6: J_{nr} = -0.0008 + 3.43 (R^2=0.48)	Rajan et al. (2023)

POTENTIAL IMPROVEMENTS TO THE MSCR TEST PROTOCOL

Despite the overall good correlation between the MSCR test and asphalt mixture rutting performance observed in these studies, there have been instances where research has indicated a weak correlation (White, 2017). In this regard, extensive endeavors have been dedicated recently to examining the potential enhancement of the MSCR test protocol, with particular emphasis on investigating the time patterns of the creep and recovery in each cycle, the adopted number of cycles, and applied stress levels.

Effect of Creep and Recovery Time

The existing test protocol with creep and recovery durations of 1 and 9 seconds, respectively, may fall short in providing a comprehensive understanding of the rutting potential of asphalt binder. Varying traffic conditions necessitate extending creep time to characterize asphalt binder behavior under heavy and slow traffic (Gaspar et al., 2019), while longer recovery time is needed for complete recovery of viscoelastic strain in modified asphalt binder (Liu et al., 2021a; Masad et al., 2009; Zoorob et al., 2012). Hence, researchers have altered the MSCR test protocol using different creep and recovery time combinations (Gaspar et al., 2019), demonstrating a significant impact on J_{nr} and recovery behavior.

For example, Kataware and Singh (2017) investigated different creep and recovery time configurations (i.e., 1/9, 2/9, 1/18, 2/18, 1/27, and 2/27 seconds). As expected, they found that extending the recovery time leads to reduced non-recoverable creep compliance and higher percent recovery compared to the outcomes from the standard MSCR test. On the other hand, increasing creep time has a reverse impact on MSCR test output. Upon increasing both creep time and recovery time at a 1:9 ratio (i.e., 1/9, 2/18, and 3/27 seconds), Gaspar et al. (2019) found that there was a higher level of non-recoverable creep compliance and a lower percentage of recovery, suggesting that the influence of increased creep time was more significant compared to the impact of extended recovery time. They also explored additional creep recovery time patterns that diverged from actual field conditions, such as 1/240 and 1/500 seconds. It became evident that, even with a 500-second recovery period, full recuperation of viscoelastic strain in a highly modified asphalt binder tested at 82°C remained unattainable. Inocente Domingas and Faxina (2022) also observed that a longer creep time revealed the negative impact of modifiers such as polyphosphoric acid (PPA), which can be used to stiffen asphalt binder.

In summary, although prolonging recovery time facilitates increased viscoelastic strain recovery and prevents overestimation of viscoplastic strain experienced during the creep phase, practicality may be a concern. In addition, finding a universally applicable recovery time for different types of asphalt binder is also challenging. Thus, Masad et al. (2009) effectively developed a mechanistic modeling approach to separate viscoplastic strain from the total strain in each cycle. The calculated J_{nr} showed a stronger correlation with the rutting potential of asphalt mixtures.

Effect of Number of Creep and Recovery Cycles

The number of creep and recovery cycles in the MSCR test has also been a subject of concern. Increasing the number of cycles can result in a more consistent strain response and better

correlations with the rutting performance of asphalt mixtures (Golalipour et al., 2017). Studies suggest that a steady-state response is typically achieved after 25–30 cycles (Behnood & Olek, 2017; Golalipour et al., 2017). However, extending the number of cycles may lead to increased damage accumulation, impacting J_{nr} and %*R*. Striking a reasonable compromise between an extended number of cycles and potential damage propagation is essential.

Effect of Adjusting and Adding Stress Levels

The standard MSCR test's two stress levels (0.1 kPa and 3.2 kPa) were chosen arbitrarily and may not represent real-life stresses for asphalt binder. In this regard, White (2017) discovered that the observed high-stress sensitivity was better accounted for by the low non-recoverable creep compliance at 0.1 kPa, rather than the high non-recoverable creep compliance at 3.2 kPa. Therefore, he suggested raising the MSCR low-stress threshold to induce sufficient deformation in the binder, thereby resulting in a non-recoverable creep compliance exceeding 0.1 kPa^{-1}. However, specifying a low stress level that is universally applicable to all types of asphalt binder across various temperatures remains a challenging task.

In response to the limitations of the standard MSCR test's stress levels, researchers have considered an additional higher stress level (e.g., 10 kPa [Golalipour et al., 2017] and 12.8–40 kPa [Jafari et al., 2022]) to enhance stress sensitivity representation. Increasing stress level in the MSCR test offers dual advantages: stronger correlation with asphalt mixture rutting performance (Jafari et al., 2022) and improved measurement repeatability compared to the lower stress level (Golalipour et al., 2017). Further exploration of the MSCR test can be found in a recent comprehensive review (Liu et al., 2021b).

J_{nrdiff}

One of the merits of the MSCR test resides in its capacity to scrutinize nonlinear behavior exhibited by asphalt binder. The assessment of stress-sensitivity relies on the calculation of the percentage difference in non-recoverable creep compliance, denoted (J_{nrdiff}) as follows:

$$J_{nrdiff} = \frac{J_{nr3.2} - J_{nr0.1}}{J_{nr0.1}} \times 100$$

Figure 8. Equation. Percentage difference in non-recoverable creep compliance.

To ensure that the $J_{nr3.2}$ value measured at the test temperature does not fall below the $J_{nr0.1}$value measured at a temperature that is 6°C higher, J_{nrdiff} is capped at a maximum of 75%. In other words, this requirement exists to ensure the stress sensitivity is not too high when grade bumping of 6°C is used. However, this indicator has prompted significant debate among researchers regarding its applicability concerning a specified threshold value. In light of the demonstrated favorable performance exhibited by a different asphalt binder that did not conform to the 75% criterion, Gaspar et al. (2019) opted to overlook said criterion when employing the MSCR test for the grading of a highly modified asphalt binder. Behnood and Olek (2017) reached a comparable conclusion, asserting that the universality of the percent difference criterion across all types of asphalt binder is disputable due to its reliance on factors such as the type of modifier and the specific testing

conditions, while Gundla et al. (2020) did not observe any correlation between this parameter and mixture rutting. To enhance the evaluation of asphalt binder stress sensitivity, Stempihar et al. (2018) recommended using a semilogarithmic graph to plot non-recoverable creep compliance against stress. Instead of relying on J_{nrdiff}, they introduced $J_{nrslope}$, which quantifies stress sensitivity by calculating the change in non-recoverable creep compliance in response to incremental stress changes (i.e., $J_{nrslope} = (J_{nr3.2} - J_{nr0.1})/3.1$). This parameter better links the alterations in non-recoverable compliance and rut depth and provided a more nuanced perspective on asphalt binder stress/temperature sensitivity. Ultimately, an extensive dataset was employed to define $J_{nrslope}$ as a function of $J_{nr3.2}$, leading to the subsequent establishment of a defined threshold.

Impact of RAP Addition on Rutting Potential of Modified Binder

Currently, many agencies use significant amounts of reclaimed asphalt pavement (RAP), for which blending the RAP with the asphalt binder and analyzing the blended residue are phases in the mix design asphalt binder selection process. In the literature, some research efforts have explored the influence of RAP addition on the MSCR outcomes of modified asphalt binders. In a study by Sing et al. (2017), the performance of crumb rubber modified asphalt binder (CRMB60 incorporated with 11% crumb rubber) blended with varying proportions (namely, 0%, 15%, 25%, and 40%) of asphalt binder extracted from RAP were assessed. Utilizing the MSCR and Linear Amplitude Sweep (LAS) methods, the rutting and fatigue resistances of CRMB60 both in the presence and absence of RAP were examined. The MSCR assessments revealed that the non-recoverable creep compliance (J_{nr}) of CRMB60 displayed a decline upon the addition of RAP up to 25%, signifying an enhancement in its resistance to rutting. Conversely, integrating a higher RAP proportion (40%) led to an elevated J_{nr} for CRMB60, signifying suboptimal rutting resistance relative to the standard CRMB60. From the LAS assessments, it was observed that the fatigue endurance of CRMB60, when combined with 15% and 25% RAP, was inferior to the baseline CRMB60.

Singh et al. (2016) evaluated the rutting potential of an SBS-modified asphalt binder (PMB40 infused with 3.5% SBS) combined with 0%, 15%, 25%, and 40% of RAP asphalt binder. They found that the inclusion of RAP asphalt binder considerably diminished the *%R* with recorded values for PMB40 combined with 0%, 15%, 25%, and 40% RAP being 89%, 55%, 17%, and 18%, respectively. The most pronounced decline was observed with a RAP composition of 25% or more. In a parallel manner, the J_{nr} metric for the PMB40 asphalt binder escalated with increasing RAP proportions, moving from 0.2 kPa^{-1} to 0.7 kPa^{-1}, 1.2 kPa^{-1}, and 1.1 kPa^{-1} as RAP increased from 0% to 15%, 25%, and 40%, correspondingly. This decline in R coupled with the spike in J_{nr} pointed toward a notably compromised resistance to rutting for the PMB40 asphalt binder when merged with RAP asphalt binder. In the graphical representation comparing *%R* and J_{nr}, data points for PMB40 blended with 25% and 40% RAP asphalt binder fell below the AASHTO benchmark. This suggests that the incorporation of the RAP asphalt binder might have adversely affected the polymer matrix of the PMB40 asphalt binder, leading to suboptimal rutting resistance.

In research by Zhou et al. (2019), PG 70-22 asphalt binder (with 3.0% SBS) was blended with RAP asphalt binder to yield blends comprising 15%, 30%, 40%, and 50% RAP asphalt binder. Evaluations utilizing the MSCR at 64°C revealed that the RAP integration led to a decrease in recovery yet had a beneficial influence on J_{nr} by reducing its value. In terms of the MSCR grading, the benchmark asphalt

binder was classified as PG64V-xx. However, blends with RAP content of 15%, 30%, 40%, and 50% were categorized as PG64V-xx, PG64V-xx, PG64E-xx, and PG64E-xx, respectively.

It can be concluded that the impact of RAP on the MSCR results of asphalt binders appears to be contingent on the modifier. Incorporating RAP in accordance with AASHTO M 332 guidelines remains a challenging endeavor, warranting additional experimental research.

ANALYSIS OF REPEATABILITY AND REPRODUCIBILITY

Extensive research has been conducted to assess the repeatability and reproducibility of the MSCR test (Zhou et al., 2014; Hossain et al., 2016). Previously, a statement on precision and bias was included in AASHTO T350-19. The estimated repeatability and reproducibility values presented in Table 3 originate from examining test outcomes of four sets of AASHTO resource proficiency samples. This analysis involved data from 149 to 225 different laboratories for each sample pair. The study focused on four modified binder grades: PG 58-28, PG 70-28, PG 76-22, and PG 82-22. In this table, 1s% denotes the coefficient of variation, while d2s% represents the acceptable range for two test results.

Precision analysis has also been provided for the ER test in Table 4. This is based on the analysis of data resulting from tests by nine laboratories, each testing three replicate specimens. Criteria for judging the acceptability of two single measurements obtained by elastic recovery test method are given as follows:

Although this test method describes a result as the average of three single measurements, the precision estimates shown as follows are based on the analysis of single measurements. For comparing two test results, the single-operator (1s) and (d2s) estimates would be reduced by a factor of $1/\sqrt{3}$. The multilaboratory (1s) and (d2s) estimates would not change.

Table 3. Estimated Repeatability and Reproducibility for MSCR Test (AASHTO T350-19)

Condition	Coefficient of Variation (1s%)	Acceptable Range of Two Test Results (d2s%)
Single-Operator Precision:		
Average Percent Recovery at 0.1 kPa, $R_{0.1}$	1.6	4.4
Average Percent Recovery at 3.2 kPa, $R_{3.2}$	1.9	5.5
Average Nonrecoverable Creep Compliance at 3.2 kPa, $J_{nr0.1}$ (kPa^{-1})	4.4	12.5
Average Nonrecoverable Creep Compliance at 3.2 kPa, $J_{nr3.2}$ (kPa^{-1})	4.7	13.2
Multi-Laboratory Precision:		
Average Percent Recovery at 0.1 kPa, $R_{0.1}$	4.8	13.5
Average Percent Recovery at 3.2 kPa, $R_{3.2}$	4.5	12.7
Average Nonrecoverable Creep Compliance at 3.2 kPa, $J_{nr0.1}$ (kPa^{-1})	11.7	33.1
Average Nonrecoverable Creep Compliance at 3.2 kPa, $J_{nr3.2}$ (kPa^{-1})	10.8	30.7

Table 4. Estimated Repeatability and Reproducibility for Elastic Recovery Test (ASTM D6084-21)

Materials Index	Standard Deviation*	Acceptable Range of Two Results*
Single-operator precision:		
1 (unmodified)	0.91	2.6
2 (modified)	0.56	1.6
Multi-laboratory precision:		
1 (unmodified)	2.32	6.5
2 (modified)	1.71	4.8

*These numbers represent, respectively, the (1s) and (d2s) limits as described in Practice C670

SPECIFICATION ADOPTION OF MSCR IN THE UNITED STATES

The current MSCR specification, AASHTO M 332, has been adopted by 16 states for all asphalt binder grades, while 25 states continue to use the *Standard Specification for Performance-Graded Asphalt Binder* (AASHTO M 320) only or employ MSCR solely as a PG Plus test. Figure 9 presents a summary of the states and their adoption of MSCR. Table 5 summarizes the states that adopted AASHTO M 332, while Table 6 summarizes the states that use MSCR as a PG Plus test but continue to use AASHTO M 320. Only three states, as far as the research team can tell, have no MSCR or ER requirement (Vermont, Mississippi, and North Carolina), now that Indiana DOT is adopting MSCR in the coming year.

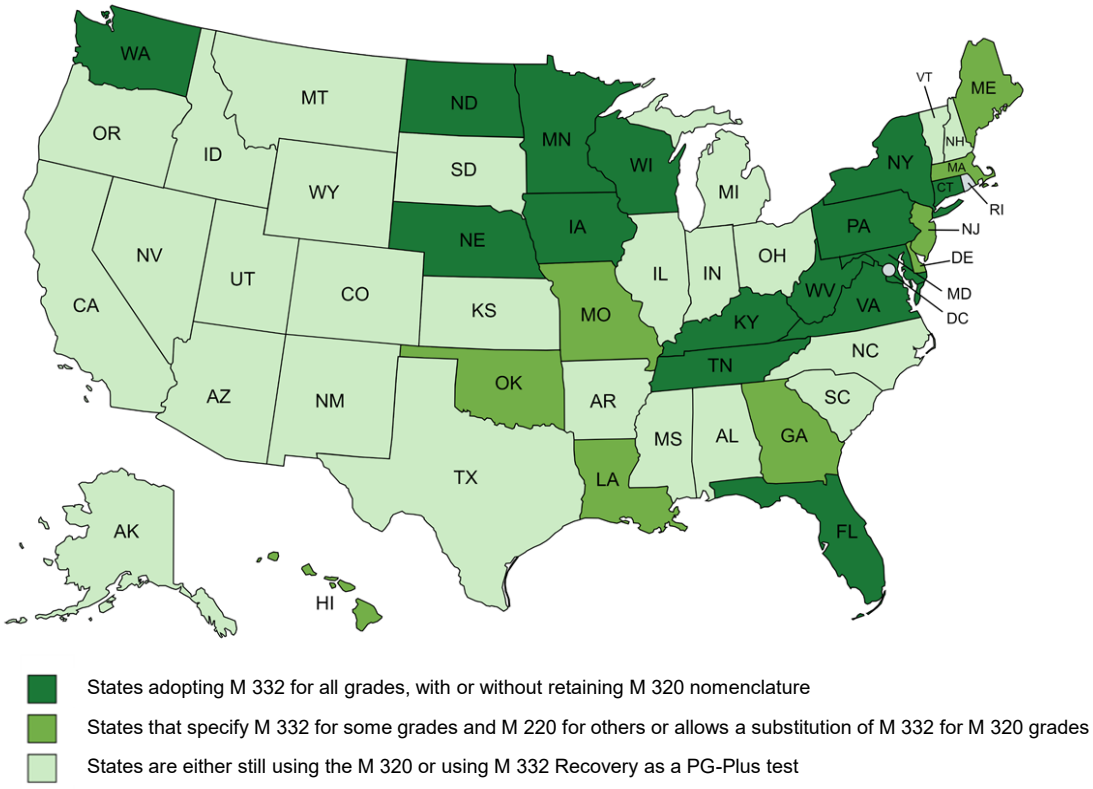

Figure 9. Graph. Summary of binder specifications (adopted from Asphalt Institute).

Table 5. Summary of Binder Grades Used in States That Adopted AASHTO M 332

State	Asphalt Binder
Connecticut	58S-28, 64S-22, 64S-28, 64E-22
Iowa	52S-34, 52H-34, 52V-34, 52E-34, 58S-28, 58H-28, 58V-28, 58E-28E
Kentucky	58-28 (58S-28), 64-22 (64-S22), 76-22 (64E-22)
Maryland	58S-22, 58S-28, 64S-22, 64S-28, 64H-22, 64E-22
Minnesota	46S-34, 52S-34, 58S-28, 58H-28, 58V-28, 58E-28, 58S-34, 58H-34, 58V-34, 58E-34, 64S-22
Nebraska	58S-34, 58H-34, 58V-34, 58E-34
New York	58E-34, 64S-22, 64H-22, 64V-22, 64E-22
North Dakota	58S-28, 58H-28, 58V-28, 58E-28, 58S-34, 58H-34, 58V-34, 58E-34
Pennsylvania	46S-40, 52S-28, 58S-28, 58E-28, 64S-22, 64S-28, 64H-22, 64E-22, 64E-28
Rhode Island	58S-28, 64S-22, 64S-28, 64E-28, 70E-34
Virginia	58S-22, 64S-22, 64H-22, 64E-22, 76E-28 (High Polymer (HP))
Washington	52S-28, 58S-22, 58H-22, 58V-22, 58S-28, 64S-28, 64H-28, 64V-28
West Virginia	58S-28, 64S-22, 64H-22, 64E-22
Wisconsin	58S-28, 58H-28, 58V-28, 58E-28, 58S-34, 58H-34, 58V-34, 58-34E
Florida	52-28, 58-22, 67-22, 76-22 (Polymer Modified Asphalt [PMA]), 76-22 (Asphalt Rubber Binder [ARB])
Tennessee	64-22, 70-22, 76-22, 82-22

Table 6. States That Have Adopted MSCR as a PG Plus Test but Use AASHTO M 320

State	Description
Alaska	J_{nr} and %*R* requirements at grade temperature for PG 52-40, 58-28, 58-34, 64-40
Nevada	J_{nr} (2.0 max) and %*R* (40% min) required at grade temperature for PG 76-22; J_{nrdiff} reported
South Carolina	J_{nr} (1.0 max) and %*R* (based on polymer modification curve in AASHTO R92) required at 64°C
Texas	%*R* requirements at grade temperature at 0.1 kPa stress level
Utah	Only reporting required

LINKING MSCR TO ELASTIC RECOVERY TEST RESULTS

Using the DSR, Clopotel and Bahia (2012) put forth a novel approach called elastic recovery in the DSR (ER-DSR), intended as an alternative to the conventional elastic recovery (ER) testing method, which employs the ductility bath as the measurement apparatus. Their results indicated that the ER-DSR could capture the effects of additives and potentially replace the conventional ER test. The test overall could predict ER results with a coefficient of correlation of 0.85. However, there has not been a wealth of studies since then relating MSCR parameters to ER results, which are still used by many agencies who have not yet implemented MSCR.

In Illinois, MSCR has not yet been implemented in any form, and the current asphalt binder specification still requires ER as a PG Plus test for polymer-modified asphalt binder. While it is understood well that the ER and MSCR results will not be a perfect 1:1 match, it is possible that MSCR could serve as a viable replacement for ER in the specification, because the goal of ER is to predict a binder's ability to recover after deformation, similar to MSCR. It should also be noted that MSCR could potentially lead to a better indication of quality and quantity of polymer modification, compared to ER which serves as a simple check for polymer. However, it is not well understood how viable prediction of ER by MSCR results is. The goal of the next chapter, therefore, is to develop a predictive method for ER based on MSCR, high PG, and other known properties. Because IDOT has collected extensive MSCR data for modified binder over the last 17 years, this is possible. If a method at least as effective as the ER-DSR method can be developed, then it is likely advantageous for Illinois to implement MSCR and consider eliminating ER due to its time- and resource-consuming nature. It should also be noted that much of the next chapter is derived directly from a recently published journal article on the topic (Asadi & Hajj, 2024).

CHAPTER 2: ANALYSIS OF EXISTING IDOT MSCR DATA

DATA EXPLORATION

The Illinois Department of Transportation (IDOT) has performed internal MSCR testing on asphalt binder from various sampling sources over the last 17 years. This study analyzed IDOT's existing MSCR dataset, along with other available high-temperature data for asphalt binder since 2006. The dataset includes MSCR non-recoverable creep compliance (J_{nr}) and percent recovery (%R) at two stress levels (i.e., 0.1 and 3.2 kPa), elastic recovery (ER) as measured by ASTM D6084 Procedure A, high performance grade (PG), and the estimated content of polymer modification derived from the results of the Fourier-transform infrared (FTIR) spectroscopy test, as well as polyphosphoric acid (PPA) estimation from x-ray fluorescence. Note that FTIR cannot accurately predict polymer content, so the values provided were only semi-quantitative. The initial dataset encompassed 1,457 data points, yet only a subset of data points included the complete set of features necessary for the prediction of the dependent variable, ER. Subsequently, through data cleansing and the application of anomaly detection techniques to identify the presence of outliers, a refined dataset comprising 703 data points was obtained. Table 7 presents the summary statistics of the dataset. Table 8 and 9 present the summary statistics for PG XX-22 and PG XX-28 asphalt binders, respectively.

Table 7. Dataset Summary Statistics for All Asphalt Binders

Variable	Mean	Standard Deviation	Skewness	Kurtosis	Min	25% Quantile	50% Quantile	75% Quantile	Max
PG-HT	71.39	3.87	-0.26	-0.7	64	70	70	76	76
%R_0.1	65.57	15.15	-0.34	-0.54	21.6	54.62	67.62	76.49	95.97
%R_3.2	61.56	19.96	-0.5	-0.71	11.69	46.07	67	76.47	96.67
J_{nr}_0.1	0.33	0.29	2.17	5.94	0.02	0.14	0.24	0.42	1.78
J_{nr}_3.2	0.41	0.42	2.41	6.93	0.01	0.14	0.26	0.52	2.63
J_{nr}_diff (%)	17.04	15.93	0.66	2.51	-43.05	6.72	14.25	25.83	95.95
Polymer (%)	3.06	1.48	-0.69	0.43	0	2.65	3.2	3.97	7.3
Elastic Recovery	85.54	5.38	-0.26	-0.46	68	81	86	89	99

Table 8. Dataset Summary Statistics for 472 PG XX-22 Asphalt Binders

Variable	Mean	Standard Deviation	Min	25% Quantile	50% Quantile	75% Quantile	Max
R_0.1	63.38	14.76	17.91	52.62	66.32	73.14	92.69
R_3.2	59.75	19.03	0.02	44.77	65.10	74.01	92.52
Jnr_0.1	0.31	0.62	0.04	0.14	0.22	0.40	13.12
Jnr_3.2	0.33	0.25	0.03	0.14	0.24	0.47	1.24
Jnr_diff	14.35	12.00	-43.05	6.51	13.21	21.43	65.26
Polymer (%)	2.96	1.38	0.00	2.60	3.08	3.78	7.30
Elastic Recovery	84.49	5.19	3.80	81.00	85.00	88.00	98.00

Table 9. Dataset Summary Statistics for 231 PG XX-28 Asphalt Binders

Variable	Mean	Standard Deviation	Min	25% Quantile	50% Quantile	75% Quantile	Max
R_0.1	68.16	18.32	0.01	55.48	72.96	81.75	96.70
R_3.2	63.13	23.93	6.01	45.04	69.33	83.50	96.67
J_{nr}_0.1	0.45	0.43	0.02	0.14	0.31	0.61	1.94
J_{nr}_3.2	0.60	0.64	0.01	0.13	0.33	0.82	2.64
J_{nr}_diff	28.47	80.95	-43.05	7.90	20.23	37.84	1210.27
Polymer (%)	3.16	1.74	0.00	2.60	3.58	4.20	7.30
Elastic Recovery	87.43	5.71	68.00	84.00	89.00	91.00	99.00

Figure 10 presents a scatterplot matrix designed to illustrate the bivariate relationships among various combinations of all variables in the dataset. The diagonal cells depict the correlation between each variable and itself using a density function. Percent recovery and J_{nr} exhibit asymmetrical distributions toward the left and right side at both stress levels, respectively. The observed trend of MSCR parameters with ER also aligns with the anticipated expectations. Moreover, based on the absence of discernible unusual patterns and outliers in the scatterplot matrix, it can be inferred that the dataset exhibits a relatively clean structure, indicating its potential suitability for the development of machine learning models. The target variable ER has shown a roughly normal distribution, and thus there is no need to employ techniques required for handling imbalanced datasets.

The Pearson product-moment correlation coefficient (r) was employed to determine the strength of the linear association among all variables. This coefficient quantifies the relationship between variables *x* and *y* by dividing the covariance ($COV(x, y)$) by the product of their respective standard deviations (σ_x and σ_y). To provide a comprehensive overview of the relationships between variables, a correlation matrix heatmap was utilized. Figure 11 presents a visualization summarizing the pairwise relationships among the variables for gaining insights into the overall correlation structure and potential associations within the dataset to be used as a basis for feature selection. Based on the empirical evidence that the polymer content lacks a predetermined parameter in real-world cases, coupled with its weak correlation with ER (r = 0.23), the research team disregarded this feature in subsequent analyses. It is noteworthy that despite the strong correlations (r = 0.97 and 0.99) between %*R* and J_{nr} at two stress levels, all of these features were retained due to the absence of overfitting concerns.

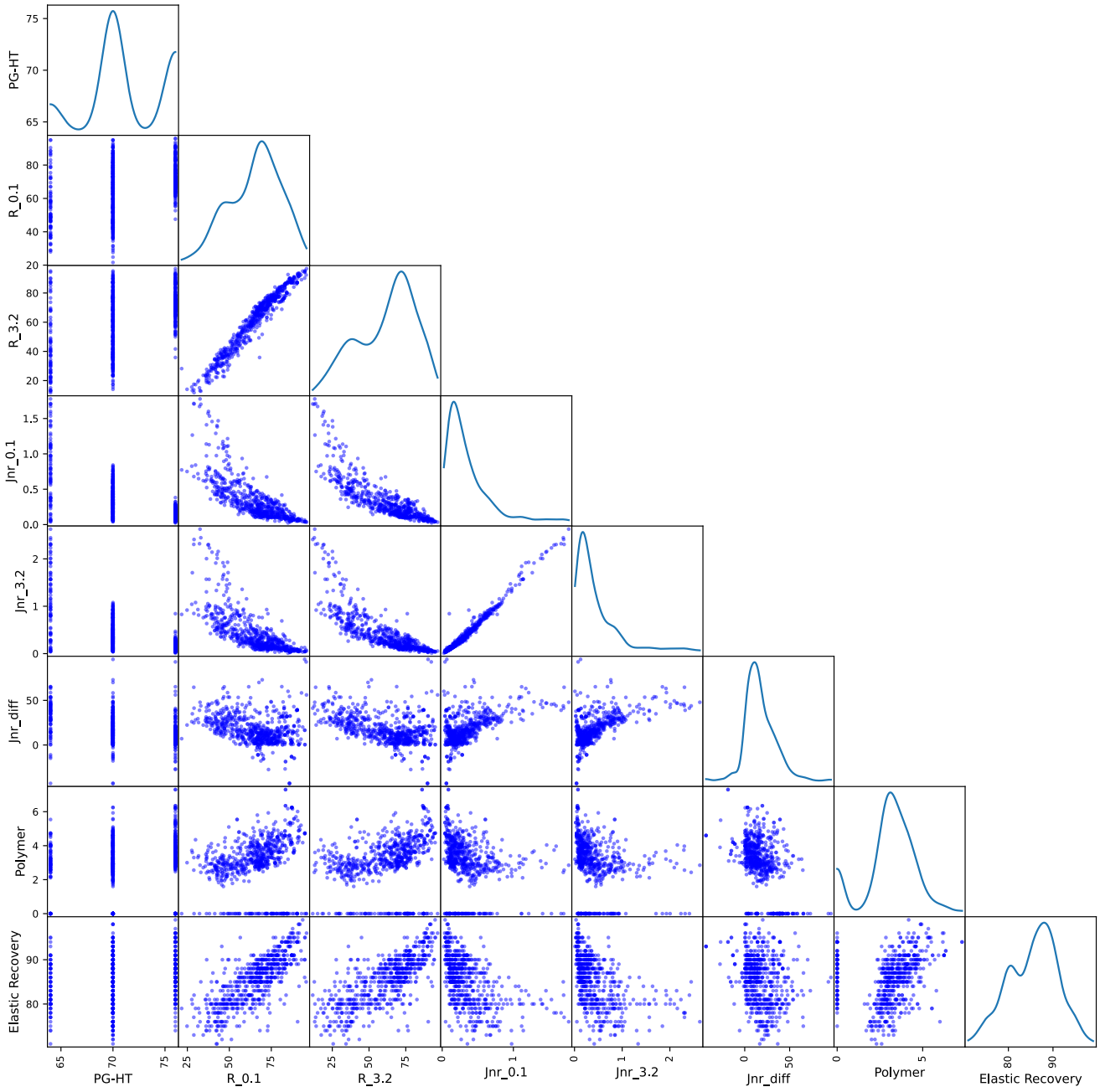

Figure 10. Graph. Visualization of pairwise relationships among variables using a scatterplot matrix.

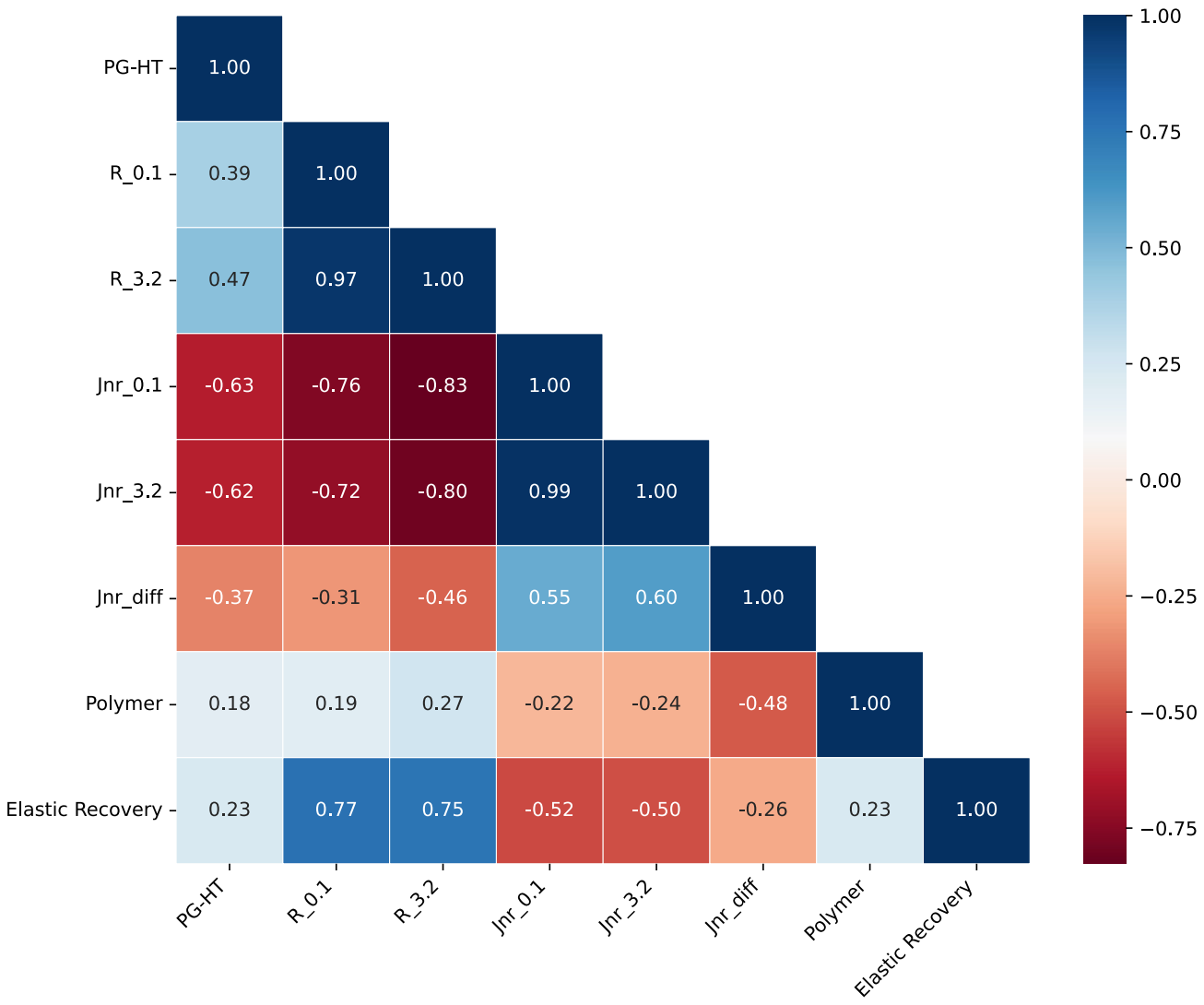

Figure 11. Graph. Heatmap of Pearson correlation matrix.

Impact of PPA Modification on MSCR Results

Because of lack of data, one factor that could not be included in the dataset discussed above was the modification of asphalt binder with polyphosphoric acid (PPA). However, IDOT did identify a limited number of asphalt binders within the data provided that were suspected to be modified with PPA, which was semi-quantitatively identified using x-ray fluorescence. As part of this research, the research team aimed to determine if computed PPA content had any specific relationship with MSCR results. Figure 12 presents an example of the data for PPA content versus %*R* for the asphalt binder in this study, with two outliers removed because they showed greater than 90% PPA content, which was assumed to be an experimental or reporting error. There is not any apparent relationship between %*R* and PPA content at either stress level, with neither high nor low recovery observed. Therefore, it is likely that more controlled experiments need to be conducted that isolate the effect of PPA modification for a specific base binder to determine its effect on MSCR parameters.

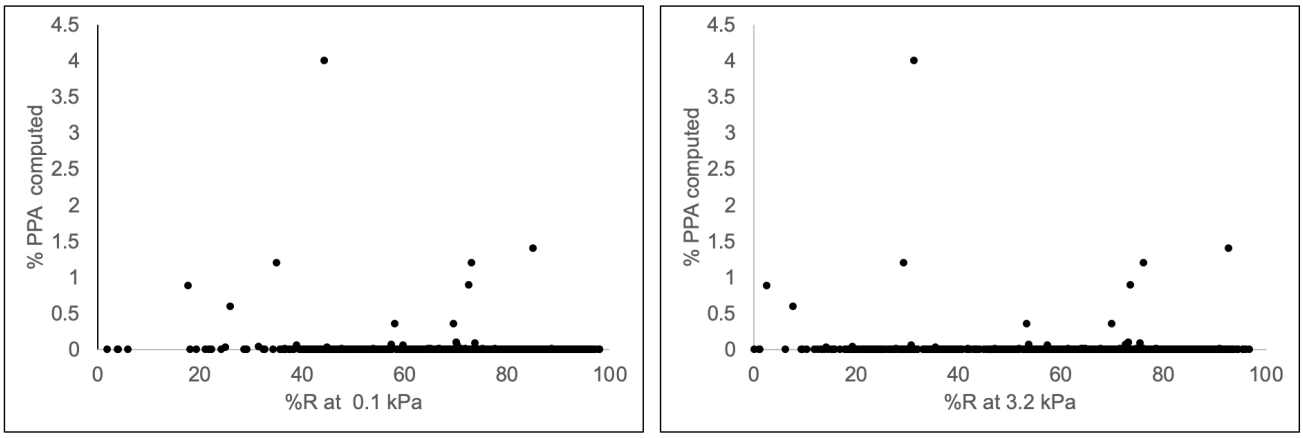

Figure 12. Graph. %*R* vs. PPA content at (a) stress level of 0.1 kPa and (b) stress level of 3.2 kPa.

% RECOVERY AS A SUPPLEMENTARY CRITERION FOR EVALUATING THE ELASTIC RESPONSE

The existing AASHTO M 332 outlines a functional correlation between J_{nr} and %*R* at 3.2 kPa, which can serve as a basis for either approving or disapproving polymer-modified binders. In the context of the MSCR test conducted on an asphalt binder, the graphical representation of percent recovery versus non-recovery creep compliance, as delineated in AASHTO R 92-18, can serve as a means to identify the presence of an elastomeric polymer based on the equation in Figure 13. If the point plotted for $R_{3.2}$ versus $J_{nr3.2}$ aligns with or surpasses the curve, it indicates that the asphalt binder has been modified with an elastomeric polymer that meets the established criteria. Conversely, if the point falls below the curve, it signifies that the asphalt binder lacks such elastomeric polymer modification. Notably, $J_{nr3.2}$ values less than 0.1 kPa^{-1} must exhibit a minimum $R_{3.2}$ value of 55%, while $J_{nr3.2}$ values exceeding 2 kPa^{-1} are not bound by any minimum $R_{3.2}$ requirement.

$$\%R_{3.2} = \begin{cases} 55 & J_{nr3.2} \leq 0.1\text{kPa}^{-1} \\ 29.371(J_{nr3.2})^{-0.2633} & J_{nr3.2} > 0.1\text{kPa}^{-1} \end{cases}$$

Figure 13. Equation. AASHTO %R versus J_{nr} relationship to identify the presence of a polymer.

The current formula in the AASHTO specification erroneously categorizes a significant portion of unaltered binders as altered ones (see the continuous black curve in Figure 14). Considering these concerns, the research team aimed to revise the AASHTO M 332 curve with the objective of establishing boundaries that would minimize the potential for any adverse impact on binder performance resulting from the modification. Note that there were also a substantial number of binders, 86 to be exact, which are modified but fell below the black line representing the current AASHTO specification. The binders above J_{nr} of 1.5 kPa^{-1} are nearly all -28 binders, and therefore the high values of J_{nr} and low *%R* can be attributed to the wrong testing temperature, as the base binder before modification was likely high PG 58 rather than 64. Therefore, the testing temperature of 64°C was potentially the wrong temperature to examine this criterion, which was developed for PG 64 base

binders if this temperature is used for testing. However, the binders that had low recovery but fell below J_{nr} of 1.5 kPa^{-1} were nearly all -22 in low PG. These binders are by all appearances modified but still failed the current criterion and proposed ones. In Figure 14, FN represents "false negative," or binders that are modified but fail, and FP represents "false positive," or binders that are unmodified but passed the curve from AASHTO M332.

Because of lack of data on unmodified binders, there were not many "false positive" binders. However, the ones that did fall just above the existing and proposed specification lines seemed to mostly be those few that were modified by ground tire rubber, which were considered "unmodified" as they did not contain modification by SBS in line with IDOT's current specification. There was also one unmodified PG 70-22 binder, which is rare for Illinois, that appeared as a "false positive." Finally, note that the proposed updated specification lines have advantages and disadvantages. Both lines are less likely to lead to false positives but will exclude more unmodified binders. However, this would not be a major issue, as it is still recommended to use ER as a backup test for failing modified binders, or to use our provided ER prediction tool. The pink line, however, could be used to comprehensively screen modified binders, even those with low PG of -28 but tested at 64°C, which provides a significant advantage in terms of implementing the test. Also, both have more stringent requirements than the current specification for low *%R* binders, which are appropriate based on the dataset.

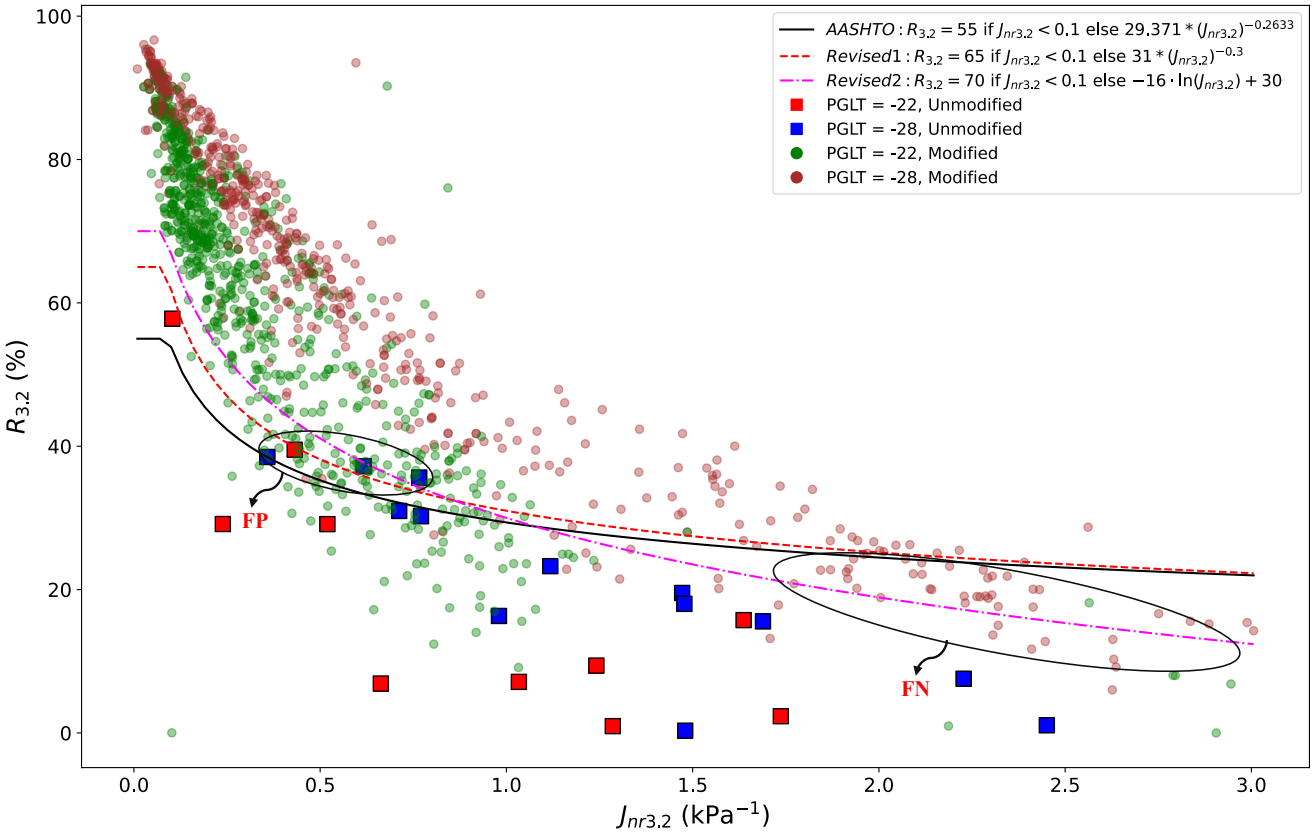

Figure 14. Graph. Prediction of polymer modification using MSCR results.

The revised limits, which are based on the comprehensive IDOT dataset employed in this research, are illustrated in Figure 14, and the associated relationships are detailed in equations in Figure 15 and 16.

$$\%R_{3.2} = \begin{cases} 65 & J_{nr3.2} \leq 0.1\text{kPa}^{-1} \\ 31(J_{nr3.2})^{-0.3} & J_{nr3.2} > 0.1\text{kPa}^{-1} \end{cases}$$

Figure 15. Equation. Revised %R versus J_{nr} relationship: first option.

$$\%R_{3.2} = \begin{cases} 70 & J_{nr3.2} \leq 0.1\text{kPa}^{-1} \\ -16ln(J_{nr3.2}) + 30 & J_{nr3.2} > 0.1\text{kPa}^{-1} \end{cases}$$

Figure 16. Equation. Revised %R versus J_{nr} relationship: second option.

MACHINE LEARNING MODEL FOR PREDICTING ELASTIC RECOVERY

Based on the above results, elastic recovery could not be predicted by a single parameter from MSCR in a highly reliable way. Overall, *%R* at both stress levels had the best relationship with ER, as partially expected, but it was evident that neither was as effective as the prior relationship observed by Clopotel and Bahia (2012). Therefore, the research team envisioned a possible use of machine learning methods for predicting ER based on multiple inputs from MSCR as well as high PG and other known information about the binders. The following section provides background on the techniques employed to achieve this.

Ensemble learning is a strategy employed to merge multiple machine learning (ML) algorithms, resulting in improved performance compared to using the algorithms individually. Instead of relying on a single model, the predictions of individual learners are combined using a combination rule to generate a more accurate and robust prediction. Ensemble learning techniques have proven highly effective in diverse machine learning applications and can be broadly classified into parallel and sequential ensembles. These approaches share similarities in aggregating the outcomes of multiple models; however, notable differences exist in the methods used to combine them. They exhibit clear distinctions in their training processes, strategies for addressing the bias-variance tradeoff, and approaches to diversifying weak learners.

Bagging

Breiman (1996) introduced bagging, which stands for bootstrap aggregating, and highlighted that perturbing the learning set could result in significant modifications in the obtained predictor. In essence, the bagging method encompasses the division of training data for each base learner through random sampling. This process generates diverse subsets that are employed to train individual base

models independently and in parallel. The purpose of bagging is to reduce variance by averaging the results from different models. This technique excels when the individual models have high variance and a tendency to overfit the training data. The primary objective of bagging is to reduce variance by combining the results obtained from multiple models. This technique particularly excels when the individual models exhibit high variance and a propensity to overfit the training data. In other words, when the algorithm used for model learning is inherently unstable (e.g., a decision tree), bagging significantly enhances the performance of base learners. Bagging delivers a valuable advantage by leveraging bootstrapping to introduce diversity in the input data, while reducing variance without increasing bias. It is worth noting that because the output of a bagging model is a combination of multiple learners, the resulting output consists of an average prediction along with a measure of variance. The latter can be interpreted as an indication of the uncertainty quantification of predictions.

Random Forest Model

The random forest algorithm (Breiman, 2001; Ho, 1995) is widely recognized as one of the most impactful bagged methods applied across a range of domains in the field of machine learning. A random forest can be seen as a collection of bagged decision trees, wherein the splitting criteria are slightly adjusted. The algorithm can be described as follows:

1. m datasets D_1, D_2, …, D_m are sampled with replacement from the original dataset D.

2. For each subset D_j, a decision tree $h_j()$ (max-depth = ∞) is trained according to the following criterion: prior to each split, a random subsampling of $k \leq d$ features (without replacement) are performed, considering only these features for the split. The default value for k in a regression problem is $\lfloor d/3 \rfloor$, and the minimum node size is 5 (Hastie et al., 2009).

3. The final prediction is: $h(\mathbf{X}) = \frac{1}{m}\sum_{j=1}^{m} h_j(\mathbf{X})$

In general, random forest employs bagging and feature randomness, specifically utilizing the random subspace technique, to construct a forest of decision trees that are uncorrelated. Furthermore, this algorithm is widely recognized as one of the most user-friendly ML models due to its minimal hyperparameter tuning demands and its ability to effectively handle a wide range of features, regardless of their type and scales, without requiring extensive preprocessing.

Extra Trees Model

Extra trees (ET) are an extension of random forest that introduce additional randomness into the tree-building process (Geurts et al., 2006). Despite their similarities, the two models are distinguishable in two key ways. First, instead of utilizing a bootstrap sample, the entirety of the learning sample is employed to train each tree in ET, potentially leading to an increase in variance. Second, the selection of cut points to split nodes are different. In ET, instead of calculating the locally optimal cut-point for each feature being considered, a random threshold is selected. The value is picked at random from a normal distribution that covers the range of the features in the training set. The node is finally divided according to the best scoring of many randomly generated splits (Geurts et

al., 2006). With the approach of randomly selecting split points, the ET algorithm also offers computational efficiency compared to random forest.

Boosting

In response to the well-known question posed by Michael Kearns about whether it is possible to combine weak learners (***H***: models slightly better than random guessing) to create a strong learner with low bias, Schapire (1999) introduced the influential boosting algorithm. These types of models are a group of ensemble models trained sequentially, with each subsequent model benefiting from the experiences of its predecessors. Starting with a simple weak learner, more complex models are developed by iteratively optimizing the weights and gradients of the training instances to reduce bias. In other words, all predictions from the weak learners are added together, with greater weight given to the models with the best track records. Due to its iterative nature, boosting is a computationally intensive process with high complexity compared to bagging models whose parallelization allows them to be computationally efficient.

An ensemble ($H_T(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$) is generated through an iterative process, wherein the regressor $\alpha_t h_t(x)$ is sequentially incorporated into the ensemble during each iteration t. The stage-wise construction of the ensemble bears a resemblance to gradient descent, with the distinction that instead of updating model parameters at each iteration, functions are added to the ensemble. Given the completion of t iterations and the existence of an ensemble $H_t(x)$, the goal in iteration $t+1$ is incorporating a new weak learner h_{t+1} to the ensemble. To this end, the weak learner that minimizes the convex and differentiable loss to the greatest extent is sought.

$$h_{t+1} = \text{argmin}_{h \in \mathcal{H}} \, \ell(H_t + \alpha h_t)$$

Figure 17. Equation. New weak learner h_{t+1} equation.

Upon finding h_{t+1}, the ensemble is updated ($H_{t+} \coloneqq H_t + \alpha h$). Now, the objective is to determine the optimal step-size α and weak learner h that minimize the loss $\ell(H + \alpha h)$. In this regard, gradient descent in the functional space is utilized, starting with the application of the Taylor approximation on $\ell(H + \alpha h)$ using the equation in Figure 18.

$$\ell(H + \alpha h) \approx \ell(H) + \alpha < \nabla \ell(H), h >$$

Figure 18. Equation. Taylor approximation of loss function.

Note that the approximation is valid in a small vicinity surrounding $\ell(H + \alpha h)$ when α is a sufficiently small constant. Using equations in Figure 17 and 18, an optimal h can be found as presented in the equation in Figure 19. Thus, boosting can be implemented when we have an algorithm $\mathbb{A}$ to solve $h_{t+1} = \text{argmin}_{h \in \mathcal{H}} \sum_{i=1}^{n} \frac{\partial \ell}{\partial [H(x_i)]} h(x)$.

$$\text{argmin}_{h\in\mathcal{H}}\ \ell(H_t + \alpha h_t) \approx \text{argmin}_{h\in\mathcal{H}} < \nabla\ell(H), h > = \text{argmin}_{h\in\mathcal{H}} \sum_{i=1}^{n} \frac{\partial \ell}{\partial [H(x_i)]} h(x_i)$$

Figure 19. Equation. Optimal solution for *h*.

The boosting models used in this study are AdaBoost (Freund et al., 1999), gradient boosting decision tree (GBDT) (Friedman, 2001), extreme gradient boosting (XGBoost) (Chen & Guestrin, 2016), light gradient boosting machine (LightGBM) (Ke et al., 2017), and categorical boosting (CatBoost) (Prokhorenkova et al., 2018).

AdaBoost

AdaBoost is the pioneering boosting model initially developed for solving classification problems, and later, it was adapted for regression tasks (Drucker, 1997). The process starts with the initialization of data sample weights. A weak learner is then trained using the modified weights obtained from the initial iteration. In each subsequent iteration, the weights of samples with higher prediction errors are increased, while the weights of samples with lower prediction errors are decreased. As the training progresses, the impact of hard-to-predict samples intensifies, causing the weak learner to focus more on previously misestimated samples. Ultimately, the final prediction is achieved through a weighted voting approach. The AdaBoost algorithm demonstrates adaptability by accommodating a wide range of base learners, with the decision tree being a frequently preferred option. Note that AdaBoost has a limitation in its susceptibility to noisy data and outliers, which arises from its iterative learning approach, potentially increasing the risk of overfitting.

GBDT

Gradient boosting, also known as gradient boosted decision tree (GBDT), is a boosting technique that creates strong ensembles by developing base learners highly correlated with the negative gradient of the loss function for the entire ensemble (Friedman, 2001). The objective of gradient boosting is to discover an approximation, denoted as $\hat{F}(x)$, for the function $F^*(x)$, that maps predictor variables x to their corresponding response variables y within the training set $S = \{x_i, y_i\}_1^N$. This is achieved by minimizing the designated loss function. The outlined procedure of GBDT can be summarized as follows (Rao et al., 2019):

Initially a constant value of mode (β) is derived:

$$F_0(x) = \underset{\gamma}{\text{argmin}}\ \ell(y_i, \beta)$$

Figure 20. Equation. Derivation of constant value of mode in GBDT.

The gradient direction of residuals is computed for each iteration, with m ranging from 1 to M, where M represents the total number of iterations.

$$\hat{y}_i = -[\frac{\partial \ell(y_i, F(x_i))}{\partial F(x_i)}]_{F(x)=F_{m-1}(x)}$$

Figure 21. Equation. Computing the gradient direction of residuals in GBDT.

The initial model $h(x_i; \alpha_m)$ is acquired by fitting the sample data, and the parameter α_m is determined using the least square method as:

$$\alpha_m = \underset{\alpha,\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (\hat{y}_i - \beta h(x_i; \alpha_m))^2$$

Figure 22. Equation. Determining α_m parameter in GBDT.

By minimizing the loss function, a new step size for the model, called the current model weight, is calculated.

$$\beta_m = \underset{\alpha,\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \ell(\hat{y}_i, F_{m-1}(x) + \beta h(x_i; \alpha_m))$$

Figure 23. Equation. Calculating the current model weight in GBDT.

Finally, the model is updated as:

$$F_m(x) = F_{m-1}(x) + \beta_m h(x_i; \alpha_m))$$

Figure 24. Equation. Updating the GBDT model.

This iterative process is performed until either the designated number of iterations is reached, or the convergence criteria are satisfied.

XGBoost

The XGBoost algorithm is a type of ensemble method that relies on decision trees and leverages the gradient boosting framework to enhance its performance. XGBoost stands apart from GBDT by including a regularization term in its objective function to prevent overfitting expressed as:

$$L_M(F(x_i)) = \sum_{i=1}^{n} \ell(\mathrm{y_i}, F(x_i)) + \sum_{m=1}^{M} \Omega(h_m)$$

Figure 25. Objective function of XGBoost.

where the second term Ω serves to penalize the complexity of the model. This regularization term is defined as:

$$\Omega(h) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2$$

Figure 26. Equation. Regularization term in its objective function of XGBoost.

in which γ is the complexity parameter that governs the minimum loss reduction gain necessary for splitting an internal node, T is the number of leaves, λ represents a penalty parameter, and w denotes the output of each leaf node. Furthermore, unlike GBDT, which utilizes the first-order derivative, XGBoost employs a second-order Taylor series expansion of the objective function. Thus, the equation in Figure 25 can be rewritten as:

$$L_M \approx \sum_{i=1}^{n}\left[g_i f_m(x_i) + \frac{1}{2}h_i f_m^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2$$

Figure 27. Equation. Second-order Taylor series expansion of the objective function of XGBoost.

where g_i and h_i represent the first and second derivatives of the loss function, respectively. The ultimate loss value is derived from the summation of all individual loss values, each of which corresponds to the respective leaf nodes in the decision tree. Given that I_j denotes all samples in leaf node j, the objective function is formulated as:

$$L_M = \sum_{j=1}^{T}[\left(\sum_{i\in I_j} g_i\right)w_j + \frac{1}{2}(\sum_{i\in I_j} h_i + \lambda)w_j^2] + \gamma T$$

Figure 28. Equation. Final form of objective function of XGBoost.

While XGBoost possesses various advantages, it has the drawback of a high number of hyperparameters, which renders the tuning process challenging.

LightGBM

The light gradient boosting machine (LightGBM) is an efficient implementation of the gradient boosting algorithm that was devised in 2017 as a versatile tool for various ML tasks (Ke et al., 2017). It leverages two innovative techniques, gradient-based one-sided sampling (GOSS) and exclusive feature bundling (EFB), to accelerate training and obtain high accuracy. The GOSS technique, a modification of gradient boosting, prioritizes instances with larger gradients, leading to faster learning and decreased model complexity. The EFB technique conducts feature selection by combining sparse and mutually exclusive attributes and then bundles these features to reduce the dimensionality of the feature matrix. Despite its advantages, LightGBM is prone to overfitting in small training datasets as well as due to the generation of more complex trees resulting from employing leaf-wise tree splitting.

CatBoost

CatBoost is a state-of-the-art gradient boosting algorithm based on decision trees (Prokhorenkova et al., 2018), renowned for its success in handling classification, regression, and ranking tasks with

categorical features. This approach combines ordered boosting, random permutations, and gradient-based optimization methodologies, facilitating remarkable performance on large and complex datasets with categorical features. It stands out from XGBoost and LightGBM due to its unique tree construction approach, which ensures balanced and symmetric trees. This strategy consistently selects the feature-split pair that minimizes loss within each iteration, leading to an optimal distribution of information across the trees. This balanced architecture reduces computational costs and serves as an effective form of regularization.

CatBoost employs an enhanced and robust approach that mitigates overfitting while ensuring all examples in the training set are effectively used for model training. This approach entails randomly permuting the training set, and for each sample, the algorithm computes the average label value of the preceding sample with the same category value in the permutation sequence. Subsequently, a specific replacement scheme is employed, incorporating a prior value and its corresponding weight. This process effectively reduces noise stemming from low-frequency categories and plays a pivotal role in enhancing the overall performance of the algorithm (Dorogush et al., 2018).

MODEL DEVELOPMENT

Initially, the dataset was partitioned into two distinct subsets, with 85% designated for training purposes and the remaining 15% allocated for the test set. Due to the inherent benefits of ensemble tree-based models in effectively managing features across varying scales, it becomes unnecessary to perform data standardization. Then, a five-fold cross-validation technique was employed to fine-tune the model hyperparameters. Typically, hyperparameter search methods consist of grid search, Bayesian optimization, heuristic search, and randomized search (Feurer & Hutter, 2019). Given its efficiency in concurrently tuning multiple hyperparameters, the randomized search method was adopted to identify the optimal hyperparameter configuration. In this regard, the training set was randomly divided into five equitably sized subsamples. Among these subsamples, one was selected as the validation set, while the other four subsamples jointly constitute the training subset. This process was iteratively repeated five times, ensuring that each subsample gets the chance to serve as the validation set once. Consequently, the average accuracy across these five validation sets is calculated to ascertain the optimal values for the hyperparameters. Subsequently, the test set was utilized to assess the performance of the models. Finally, the model with the best overall performance was chosen for verifying trustworthiness and investigating feature importance.

MODEL PERFORMANCE EVALUATION

In this study, the performance evaluation of the developed ensemble models incorporates three goodness-of-fit measures. These measures include the coefficient of determination (R^2) (the equation in Figure 29), which quantifies the portion of variance explained by the model, the mean absolute error (MAE) (the equation in Figure 30), describing the average discrepancy between measured and predicted ER values, and the root mean squared error (RMSE) (the equation in Figure 31), which describes the standard deviation of the prediction error.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{\sum_{i=1}^{n}(y_i - \bar{y})}$$

Figure 29. Equation. Coefficient of determination equation.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Figure 30. Equation. Mean absolute error equation.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Figure 31. Equation. Root mean squared error equation.

in which y_i and $\hat{y}_i$ represent the ith measured and predicted ER, and $\bar{y}$ denotes the mean of y_i values.

MODEL INTERPRETATION USING SHAPLEY VALUE ANALYSIS

Lundberg and Lee (2017) introduced Shapley Additive Explanations (SHAP) as a groundbreaking technique in the field of machine learning interpretation. This state-of-the-art approach, rooted in coalitional game theory, has gained significant traction due to its ability to unravel the intricate workings of ML models. Employing SHAP allows for a comprehensive understanding of the marginal contribution brought about by each individual feature in the prediction process. Unlike model-specific interpretation methods, SHAP is considered model-agnostic, meaning it can be applied to a wide range of models without relying on their internal architecture. At its core, SHAP operates by quantifying the impact of including or excluding a specific feature within a coalition of features. The Shapley value of a feature represents the average change in prediction when that feature joins the existing set of features. By computing these values for all features, SHAP uncovers the marginal contribution of each feature on the model's output. One of the remarkable advantages of SHAP is its ability to provide both local and global interpretability. On a local level, SHAP values are computed for individual data points, shedding light on how each feature contributes to a specific prediction. On a global scale, SHAP values provide a holistic understanding of feature importance across the entire dataset. These values not only quantify the overall impact of a feature, but also indicate whether its effect on the model's output is positive or negative. This global interpretability empowers practitioners to identify crucial features driving the model's predictions and assess the robustness of their decision-making process. Thus, the resulting SHAP values provide a fair allocation of credit to each feature, considering their individual and collective contributions to the prediction process.

RESULTS AND DISCUSSIONS

Performance of Ensemble Models

The process of tuning model hyperparameters holds critical importance in the establishment of a robust model. This step ensures the model's performance is finely calibrated, enhancing its reliability and efficacy. Table 10 and 11 provide documentation of the results derived from the rigorous optimization process, encompassing specific hyperparameter values in bagging and boosting algorithms, respectively.

Table 10. Results of Hyperparameter Optimization for Bagging Models

Algorithm	Hyperparameter	Description	Search	Selected Value
BR	n_estimators	Number of trees	[50-1000]	1000
	max_samples	Maximum number of samples for training each tree	[0.1-1]	0.3
RF	n_estimators	Number of trees	[50-1000]	400
	max_depth	Maximum depth of the tree	[5-50]	15
	min_samples_split	Minimum number of samples for nodes split	[2-20]	2
	min_samples_leaf	Minimum number of samples for leaf nodes	[1-15]	2
ET	n_estimators	Number of trees	[50-1000]	500
	max_depth	Maximum depth of the tree	[5-50]	50
	min_samples_split	Minimum number of samples for nodes split	[2-20]	10
	min_samples_leaf	Minimum number of samples for leaf nodes	[1-15]	3

Table 11. Results of Hyperparameter Optimization for Boosting Models

Algorithm	Hyperparameter	Description	Search	Selected Value
AdaBoost	n_estimators	Number of trees	[50-1000]	200
	learning_rate	Shrinkage coefficient of each tree	[0.01-0.5]	0.01
GBDT	n_estimators	Number of trees	[50-1000]	750
	learning_rate	Shrinkage coefficient of each tree	[0.01-0.5]	0.01
	max_depth	Maximum depth of a tree	[3-11]	3
	min_samples_leaf	Minimum number of samples for leaf nodes	[1-15]	4
	min_samples_split	Minimum number of samples for nodes split	[2-20]	5
XGBoost	n_estimators	Number of trees	[100-1000]	600
	learning_rate	Shrinkage coefficient of each tree	[0.01-0.3]	0.01
	max_depth	Maximum depth of a tree	[3-11]	7
	subsample	Fraction of samples used for training each tree	[0.7-1]	0.8
	colsample_bytree	Fraction of features used for tree construction	[0.6-1]	0.6
	reg_alpha	L1 regularization	[0-1]	0
	reg_lambda	L2 regularization	[0-1]	0.001
LightGBM	n_estimators	Number of trees	[100-1000]	200
	learning_rate	Shrinkage coefficient of each tree	[0.01-0.3]	0.1
	max_depth	Maximum depth of a tree	[3-11]	5
	num_leaves	Number of leaves for each tree	[10-200]	100
	bagging_fraction	Fraction of samples used for training each tree	[0.7-1]	0.8
	feature_fraction	Fraction of features used for tree construction	[0.6-1]	1
	min_data_in_leaf	Minimum number of samples to form a leaf	[10-200]	50
	lambda_l1 (or reg_alpha	L1 regularization	[0-1]	0.01
	lambda_l2 (or reg_lambda)	L2 regularization	[0-1]	0.2
	min_gain_to_split	Minimum gain to make a partition on a leaf node	[0-0.3]	0.3

Algorithm	Hyperparameter	Description	Search	Selected Value
CatBoost	iterations	Number of trees	[100-1000]	400
	learning_rate	Shrinkage coefficient of each tree	[0.01-0.3]	0.01
	depth	Maximum depth of the trees	[3-11]	9
	subsample	Fraction of samples for training each level of a tree	[0.7-1]	0.8
	colsample_bylevel	Fraction of features used for tree construction	[0.6-1]	1
	min_data_in_leaf	Minimum number of samples to form a leaf	[10-100]	10
	l2_leaf_reg	L2 regularization	[0-10]	1

Utilizing the optimal set of hyperparameters, Figures 32 and 33 display a performance evaluation of the developed bagging and boosting models, presenting a comparison between the predicted and measured ER values for the test set. As shown in Figure 29, the bagging models underwent scrutiny under varying error ranges: < 5%, 5%–10%, and > 10%. Extra trees displayed superior performance, achieving 86.36% within the < 5% range, 12.73% within 5%–10%, and 0.91% exceeding the > 10% threshold. In contrast, bagging regressor demonstrated higher error rates within the 5%–10% range, with 20% of predictions falling in this interval, and 1.82% of predictions beyond the > 10% boundary. The values of different error measures for both the training and testing sets of the bagging models are shown in Table 12. Among the evaluated bagging models, the extra trees (ET) model distinguishes itself with an extraordinary R^2 value of 0.999 during the training phase, indicating an exceptionally precise fit to the training data. On the testing data, the extra trees model maintains its prominence by achieving the highest R^2 value of 0.851, along with the lowest MAE and RMSE values of 1.64 and 2.39, respectively. These results underscore the superior predictive accuracy and generalization capacity of the extra trees model compared to the other two models. It is pertinent to note that through the manipulation of hyperparameters, we could alleviate overfitting. However, this led to diminished performance on the training set, subsequently resulting in a corresponding decline in the testing set performance. This discrepancy runs counter to our principal aim of enhancing the accuracy specifically for the testing data. For instance, the ET model's hyperparameters can be fine-tuned to mitigate the issue of overfitting, resulting in an R^2 of 0.895 for the training dataset and 0.822 for the testing dataset. However, an imperative remains to enhance the model's performance concerning the testing data; thus, the set of hyperparameters in Table 10 that yields an R^2 of 0.847 was retained. This approach has been similarly adopted for all bagging and boosting models.

Table 12. Performance of Bagging Models on Training and Testing Data

Model	Training Data			Testing Data		
	R^2	MAE	RMSE	R^2	MAE	RMSE
Bagging	0.921	1.01	1.52	0.739	2.19	2.92
Random Forest	0.943	0.91	1.27	0.814	1.83	2.58
Extra Trees	0.999	0.001	0.008	0.851	1.64	2.39

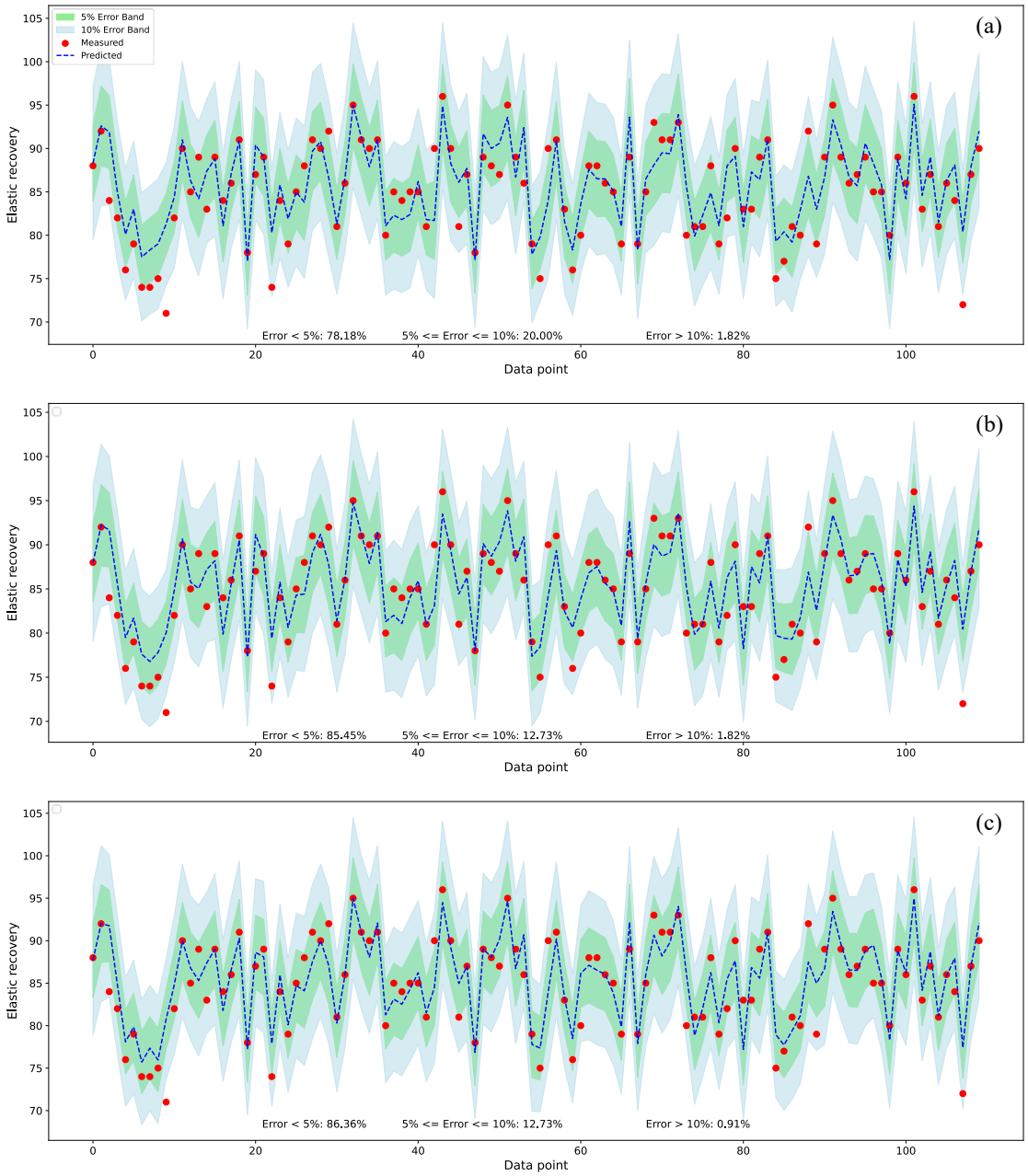

Figure 32. Graph. Performance of (a) bagging, (b) random forest, and (c) extra trees on testing data.

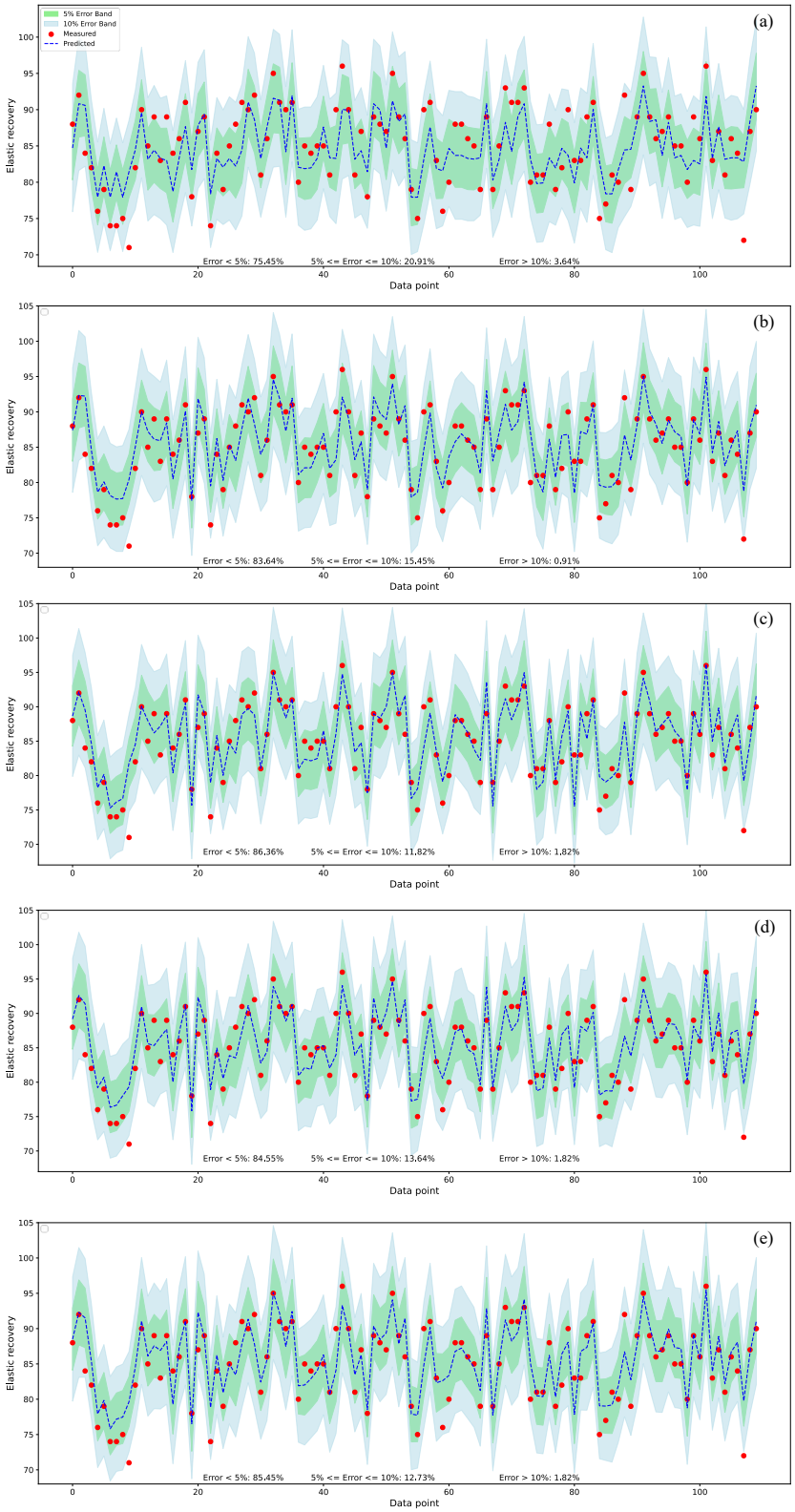

Figure 33. Graph. Performance of (a) AdaBoost (b) GBDT, (c) XGBoost, (d) LightGBM, and (e) CatBoost on testing data.

For the boosting models, the performance analysis reveals intriguing insights into their predictive capabilities (Figure 33). Table 13 provides an overview of performance criteria for both the training and testing datasets. Among the developed boosting models, XGBoost and CatBoost stand out with noteworthy R^2 values of 0.842 and 0.826 on the testing data, respectively. In addition, XGBoost achieved the lowest MAE and RMSE values of 1.69 and 2.45, respectively, signifying its precision in predicting ER values. During the training phase, XGBoost exhibited a remarkable R^2 value of 0.997, while CatBoost and LightGBM demonstrated values of 0.939 and 0.918, respectively. The relatively smaller difference between their training and testing performance suggests a reduced likelihood of overfitting for these two models. GBDT delivered acceptable performance on the testing data, whereas AdaBoost struggled to discern the inherent distribution within the training data, leading to unsatisfactory performance. The error analysis of the boosting models presented in Figure 33 revealed their distinct strengths in handling different error ranges. XGBoost demonstrated impressive precision, with 86.36% predictions within < 5% error and 11.82% predictions falling between 5% and 10% error. Meanwhile, CatBoost and LightGBM achieved competitive performance. For larger errors exceeding 10%, all superior boosting models (i.e., XGBoost, CatBoost, and LightGBM) exhibited a similar proportion of 1.82% predictions. These results shed light on the models' efficacy in handling varying degrees of prediction errors, aiding in making informed decisions for real-world applications.

Table 13. Performance of Boosting Models on Training and Testing Data

Model	Training Data			Testing Data		
	R^2	MAE	RMSE	R^2	MAE	RMSE
AdaBoost	0.627	2.73	3.31	0.613	2.79	3.48
GBDT	0.824	1.77	2.32	0.784	1.97	2.72
XGBoost	0.997	0.19	0.27	0.842	1.69	2.45
LightGBM	0.918	1.14	1.52	0.819	1.82	2.55
CatBoost	0.939	1.08	1.34	0.826	1.77	2.52

To further analyze the performance of the developed models, an examination of their residual values was conducted, as illustrated in Figure 34. The distribution of residuals ($y_i - \hat{y}_i$) for a well-performing model should adhere to a normal distribution, and the average of the residuals should be approximately zero. The examination of residuals for all models revealed a general adherence to a roughly normal distribution. However, the extra trees model demonstrated superior performance compared to the other two bagging models, as evidenced by a pronounced clustering of data points around zero residuals in Figure 34-c. Within the domain of boosting models, the residual distribution for GBDT tends to approximate the normal distribution more closely (Figure 34-e), while a greater concentration of data points is evident around zero residuals for XGBoost and CatBoost as in Figure 34-f and Figure 34-h, respectively.

It is worth noting that a one-by-one comparison between residual distributions can serve as the basis for hypothesis testing (e.g., Student's t-test or Welch's t-test for this problem) to determine whether the two models are significantly different. To be more specific, the distribution of err(M1) - err(M2)

with a mean of zero can be considered a practical null hypothesis. Then, by calculating the t-statistic: $(t = \frac{\overline{err}(M_1) - \overline{err}(M_2)}{\sqrt{var(M_1 - M_2)/k}})$ and comparing it to t0 a chosen significance level, we can investigate whether the null hypothesis is rejected. In light of the considerable number of models developed, necessitating extensive testing, and the fact that there is minimal differentiation among models with similar characteristics, the selection of superior models based on performance evaluation is deemed satisfactory.

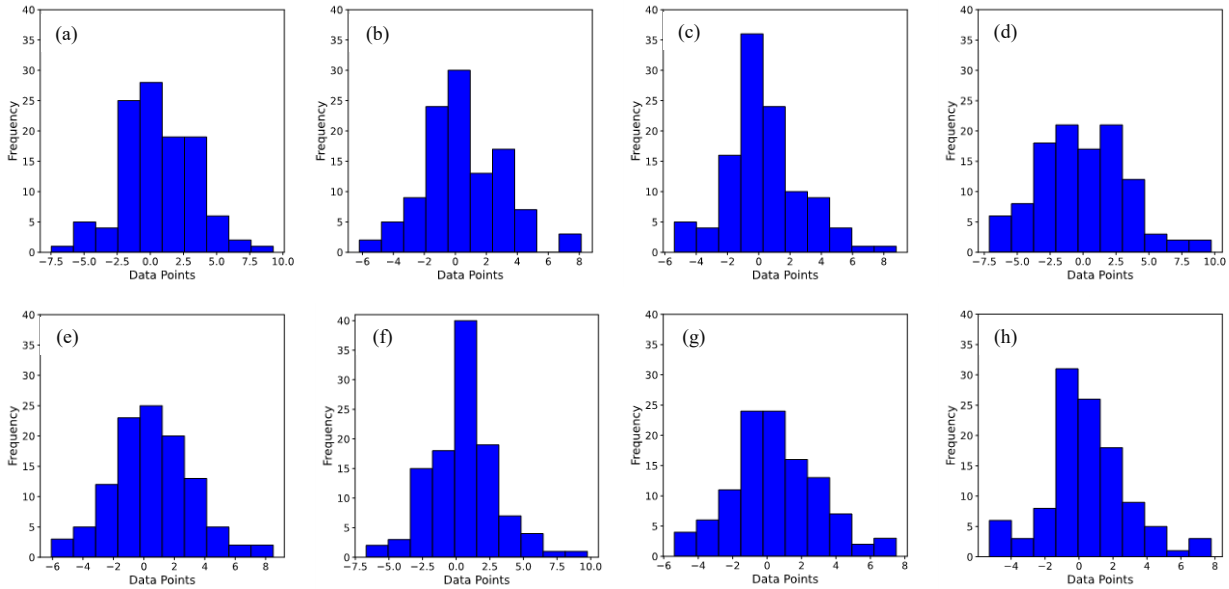

Figure 34. Graph. Residual distribution for ER predictions on the testing data for a) bagging, (b) random forest, (c) extra trees, (d) AdaBoost (e) GBDT, (f) XGBoost, (g) LightGBM, and (h) CatBoost.

The model evaluation shows that bagging and boosting methods are effective in capturing complex relationships in the dataset. Extra trees excelled among bagging models, while XGBoost emerged as the top performer among boosting models. This highlights the potential of ensemble learning approaches for robust predictions of ER from DSR high-temperature results, which are even more efficient than ER-DSR (Clopotel & Bahia, 2012). To be more specific, the correlation between ER-DSR and traditional ER was investigated among 20 types of asphalt binder, including SBS-modified PG 70-22 and PG 76-22, from 10 sources (Morshed et al., 2020). The R^2 values were determined to be 0.67 and 0.845 for unaged and RTFO-aged conditions, respectively, demonstrating the superior performance of the ET and XGBoost models developed in this study. Although it should be noted that, despite the rigorous hyperparameter tuning and utilization of ensemble learning methods to mitigate overfitting, a discrepancy is still observed between the model performance on training and testing data. Hyperparameter tuning involved fine-tuning critical parameters, including learning rates, regularization strengths, and tree depths, to strike an optimal balance between model complexity and generalization. Additionally, ensemble learning techniques (i.e., bagging and boosting) have been an integral part of our approach. By combining predictions from multiple base models, ensemble methods help in reducing the risk of overfitting. We engineered ensembles with diverse model architectures to bolster the stability and robustness of our predictions. In essence, the persistence of

overfitting can be primarily attributed to the inherent complexities and unseen noise embedded within the dataset. To be more specific, the dataset spans over a 17-year period, introducing a potential source of aleatoric uncertainty. Aleatoric uncertainties, arising from the inherent randomness in the physical attributes of the system, the stochastic nature of input excitations, and considerable noise in the experimental data (Olivier et al., 2021), contribute to fluctuations in ER measurements. Factors influencing these fluctuations encompass testing equipment, specimen preparation techniques, and potential operator influence. Consequently, further research to refine and enhance the dataset may be necessary to achieve better generalization performance on the testing data.

Comparison with Neural Networks

In addition to ensemble models, we investigated the capabilities of neural networks (NNs) and compared the model's predictive capability to those of extra trees and XGBoost. To tackle the sensitivity of neural net to differences in feature scales, all features were first normalized such that they fell within the range of zero to one. The NN's hyperparameters encompass the number of hidden layers and nodes per layer, the type of nonlinearity, the loss function, the regularization method, and parameters related to optimization including the adopted method, learning rate schedule, minibatch size, and the number of epochs.

After conducting an exhaustive search to identify the most accurate model, an architecture comprising three hidden layers, each with 50 nodes, was designed. As the type of nonlinearity, the rectified linear unit activation function was selected. To combat overfitting, regularization was employed through the dropout method, where some nodes in each forward pass during training were deactivated with a probability p of 0.15 obtained via a greedy search from p = 0.1 to 0.5. For optimization, the adaptive moment estimation (Adam) method was used. The learning rate and minibatch size were fine-tuned through a grid search, ranging from 0.1 to 1e-4 and 100 to 500, respectively. Ultimately, the minibatch size was set to 300 and the learning rate to 0.001 to achieve optimal performance.

The results of the neural network with tuned hyperparameters on the testing data are illustrated in Figure 35. In terms of the chosen model evaluation criteria, the neural network with R^2 = 0.82, MAE = 1.80, and RMSE = 2.54 performed on par with random forest and LightGBM but did not outperform extra trees and XGBoost. Nonetheless, it is noteworthy that although a small portion (0.91%–1.82%) of predictions in all ensemble models surpassed the >10% error threshold, no such instances were observed for the NN model. In conclusion, based on these findings, the NN model does not exhibit superior performance compared to the best ensemble models in this study.

The superiority of tree-based models in this study predominantly stems from the dataset's characteristics. The relatively modest tabular dataset, comprising around 700 samples and six features, may not provide the necessary complexity or patterns for a deep neural network to effectively leverage its feature extraction capabilities. Conversely, tree-based models often exhibit robustness when confronted with smaller datasets and can efficiently exploit available information. In the context of deep learning performance, it is important to note that despite its achievements in text and image datasets, its superiority on tabular data remains uncertain (Grinsztajn et al., 2022).

Recent extensive benchmarking across various datasets has illustrated the excellence of tree-based models like XGBoost and Random Forests on medium-sized data (~10,000 samples), even without considering their faster computation (Grinsztajn et al., 2022).

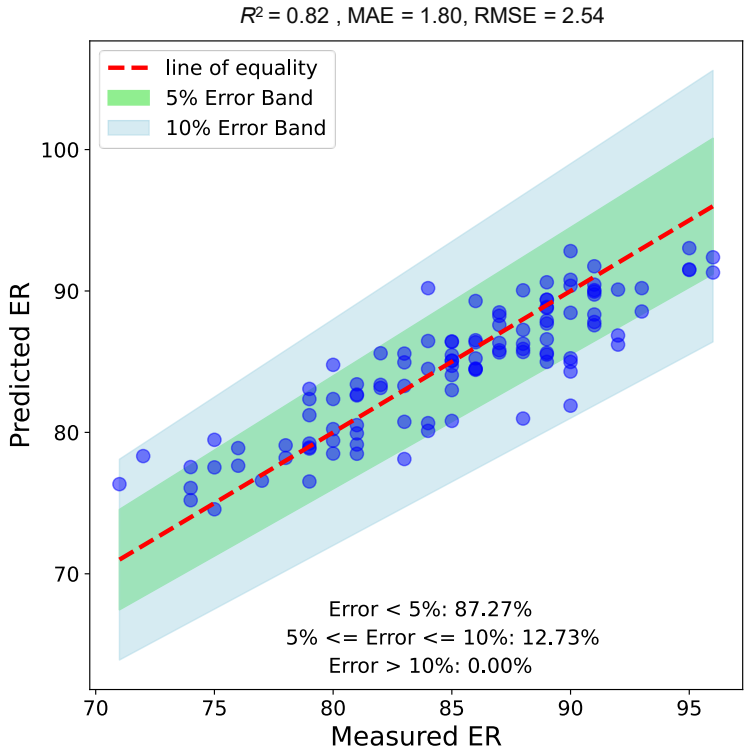

Figure 35. Graph. Performance of NN model on testing data.

Therefore, upon confirming the trustworthiness of ET and XGBoost via Shapley analysis, as presented in the next section, they were selected as the best models. It is worth noting that the models were developed utilizing the IDOT database, where most binders underwent modification with SBS being the predominant modifier. Consequently, the database exhibits relatively high bounds for elastic recovery (ER) ranging from 61 to 99, which could be a limitation of the model. Future work is expected to incorporate asphalt binder with a wider range of elastic recovery values, including straight-run, unmodified binder.

Model Interpretation with SHAP Values

In this study, we employ SHAP analysis to assess the feature importance and trustworthiness of superior ensemble models (i.e., extra trees and XGBoost models). As shown in Figure 36, both models revealed $R_{0.1}$, $R_{3.2}$, $J_{nr0.1}$, and $J_{nr3.2}$ as their top four influential features. However, their least significant features differed; for ET, PG_HT and J_{nrdiff} were the least important, while for XGBoost, J_{nrdiff} and PG_HT were identified as the least impactful in that order. Notably, a comparative examination of the relative difference between $R_{0.1}$ and $R_{3.2}$ in XGBoost and extra trees demonstrated a higher disparity in the former, indicating a more pronounced impact of $R_{0.1}$ in the latter model. The outcomes of this analysis provide essential insights into the primary determinants influencing the models' predictions.

It becomes evident that %*R* holds substantial importance in both models, suggesting its significance as a foundational element for a future MSCR specification.

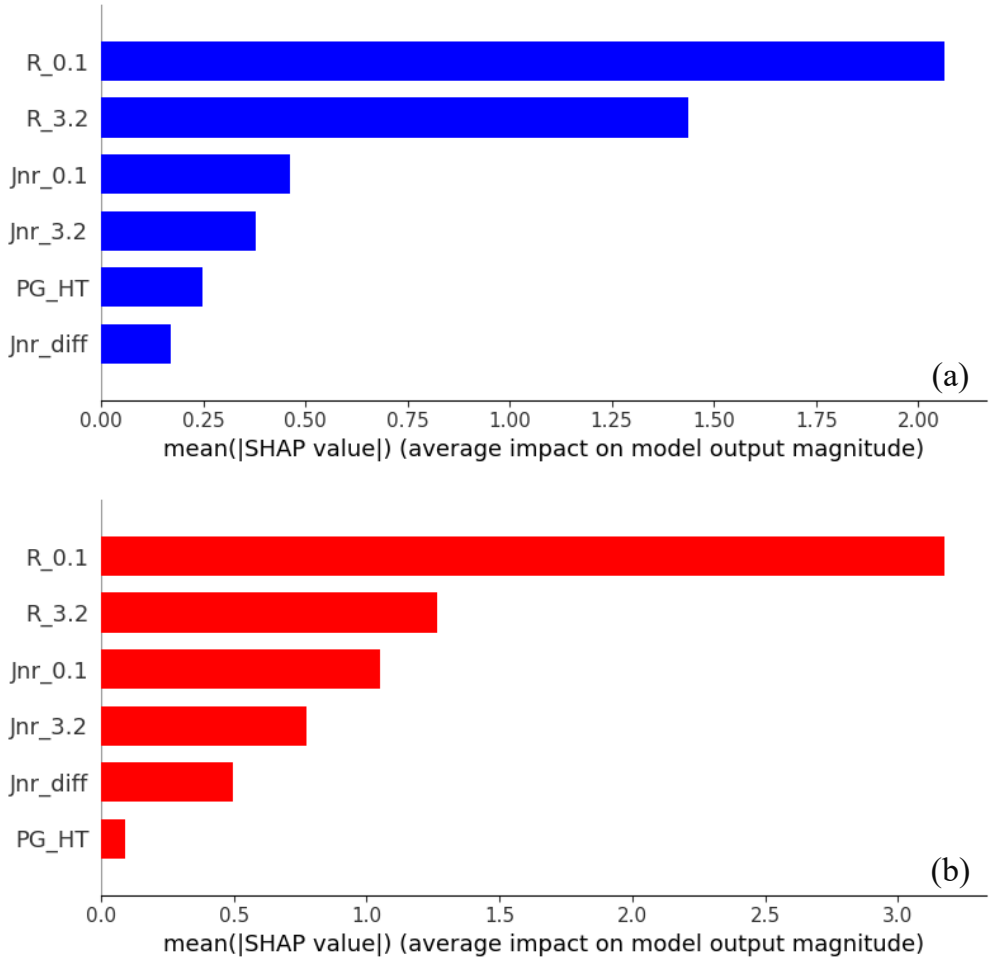

Figure 36. Graph. Feature importance for (a) extra trees and (b) XGBoost using testing data.

Figure 37 presents an analysis of the impact of dataset features on elastic recovery (ER) as the output of the developed models. The x-axis portrays the SHAP value assigned to each test point, serving as an indicator of the feature's relative significance for that specific instance. Each row in the visualization corresponds to a distinct feature, where the color of the points conveys the magnitude of the feature, with red denoting higher values and blue indicating lower values. In accordance with physical expectations, both models captured a positive correlation between %*R* for both stress levels. Conversely, the models demonstrate a negative correlation between J_{nr}, where the trend is more pronounced at the stress level of 3.2 kPa. Generally speaking, the results obtained from the SHAP analysis are consistent with prior research (D'Angelo, 2010; Moraes et al., 2017), providing supporting evidence that the MSCR test has the potential to replace the ER test, with a major focus on the %*R* parameter. In addition, the lower impact of the percent difference in non-recoverable creep compliance (J_{nrdiff}) is in line with previous studies that have shown the limitation of this parameter with a 75% threshold (Behnood & Olek, 2017; Gaspar et al., 2019; Stempihar et al., 2018).

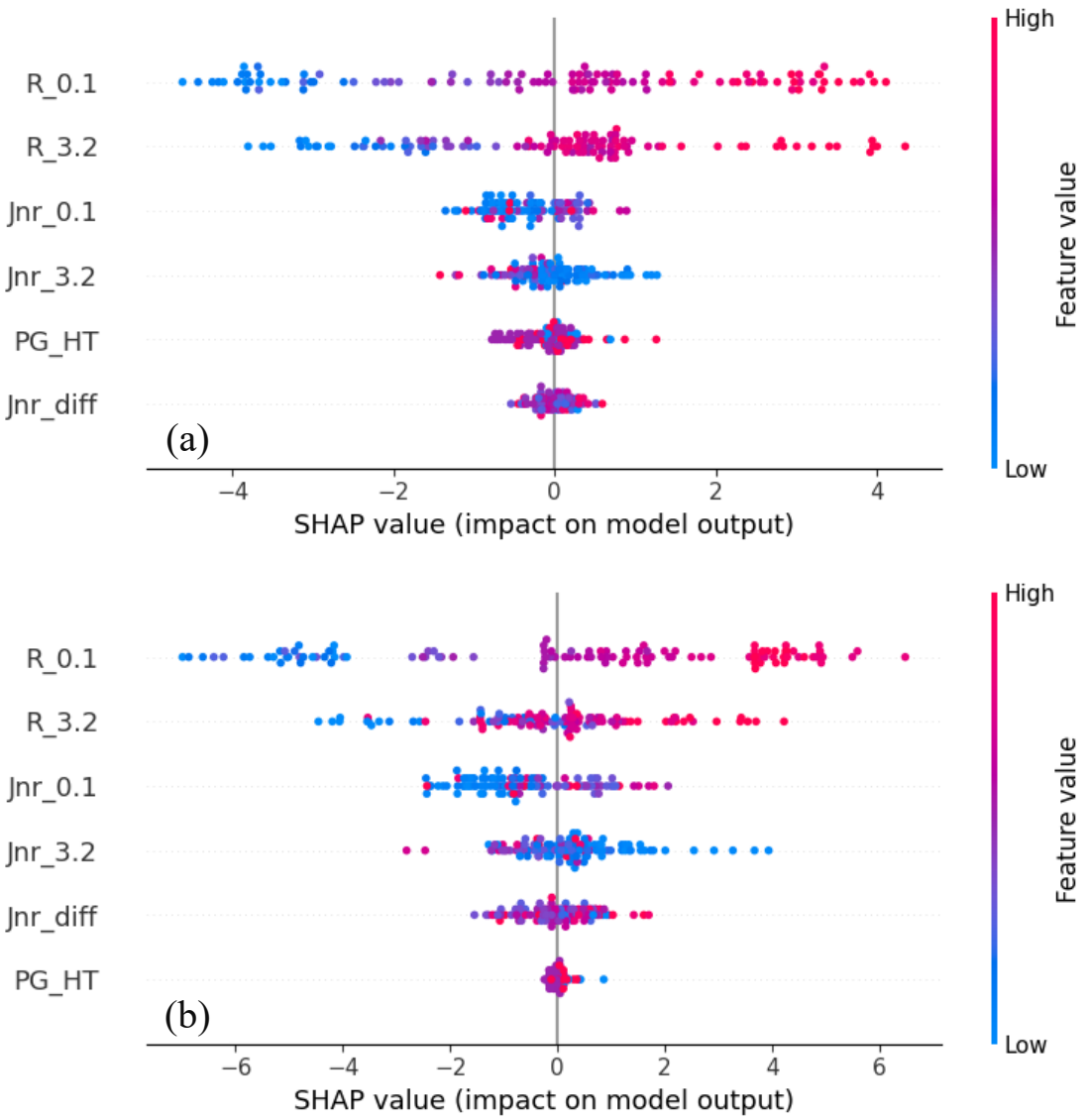

Figure 37. Graph. Model interpretation for (a) extra trees and (b) XGBoost.

Binder Clustering Using MSCR Results

In this section, clustering analysis was employed to study asphalt binder based on MSCR parameters. Through the application of clustering techniques, the research team aimed to reveal intrinsic patterns and differences within the database, shedding light on how various types of binder can be distinguished using their MSCR characteristics. To this end, we first employed principal component analysis (PCA) to reduce the dimensionality of the dataset. Subsequently, the researchers determined the optimal number of clusters, and then applied the K-means algorithm to partition the data into the identified clusters to reflect their inherent similarities. PCA is a dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional subspace while preserving the most important information. In this context, the first and second principal components explained 76.1% and 23.3% of the dataset's variance, capturing a comprehensive picture of the data distribution. The

next critical step is to ascertain the optimal number of clusters within the dataset. The elbow method and Silhouette score are among the most common approaches in estimating the appropriate number of clusters (Nanjundan et al., 2019). The silhouette score evaluates how well a data point fits within its own cluster. Its value can range from −1 to 1, with higher values indicating better cluster coherence, while negative values suggest that data points may have been clustered incorrectly. The underlying concept of the elbow method is to ascertain the juncture in the plot where the inertia begins to flatten and create a distinct elbow shape. This critical point signifies that further inclusion of clusters no longer substantially reduces inertia, and it can be considered as the optimal choice of cluster numbers.

Figure 38 demonstrates the results obtained by employing the Silhouette score and elbow method on the dataset. As depicted in Figure 38-a, the optimal number of clusters with the highest Silhouette score was two. The plot of the elbow method offers valuable indications of the optimal cluster number, although its determination remains subjective. The observed possibilities for the optimal cluster count are 2, 3, or 4, with 2 clusters aligning with the result derived from the Silhouette score.

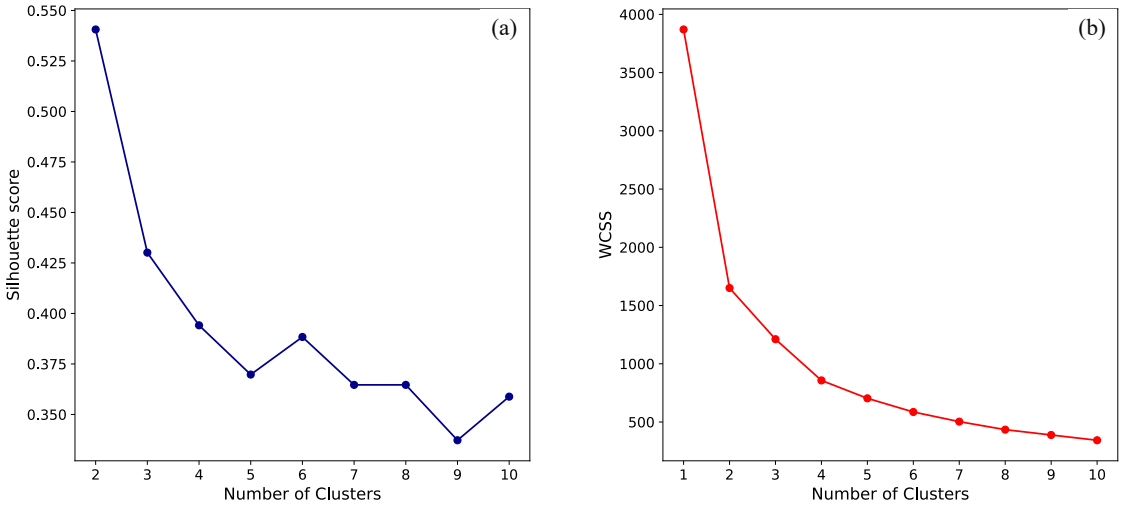

Figure 38. Graph. Determining the optimum number of clusters using (a) Silhouette method and (b) elbow method.

Continuing the cluster analysis, the researchers proceed to visualize the dataset's clustering outcomes based on different cluster configurations using the first and second principal components, as shown in Figure 39. The discernment of distinct patterns within clustered binders poses a considerable challenge, particularly when considering the knowledge of six different performance grades (i.e., PG 64-28, PG 64-22, PG 70-28, PG 70-22, PG 76-28, and PG 76-22), exhibiting polymer content within the range of 0%–7.3% in the dataset. Notwithstanding the overall efficacy demonstrated by the MSCR test, there exist promising avenues for the development of novel testing methodologies or the derivation of additional parameters from the MSCR test. Such endeavors are intended to enhance the differentiation capability among the diverse types of asphalt binder more effectively.

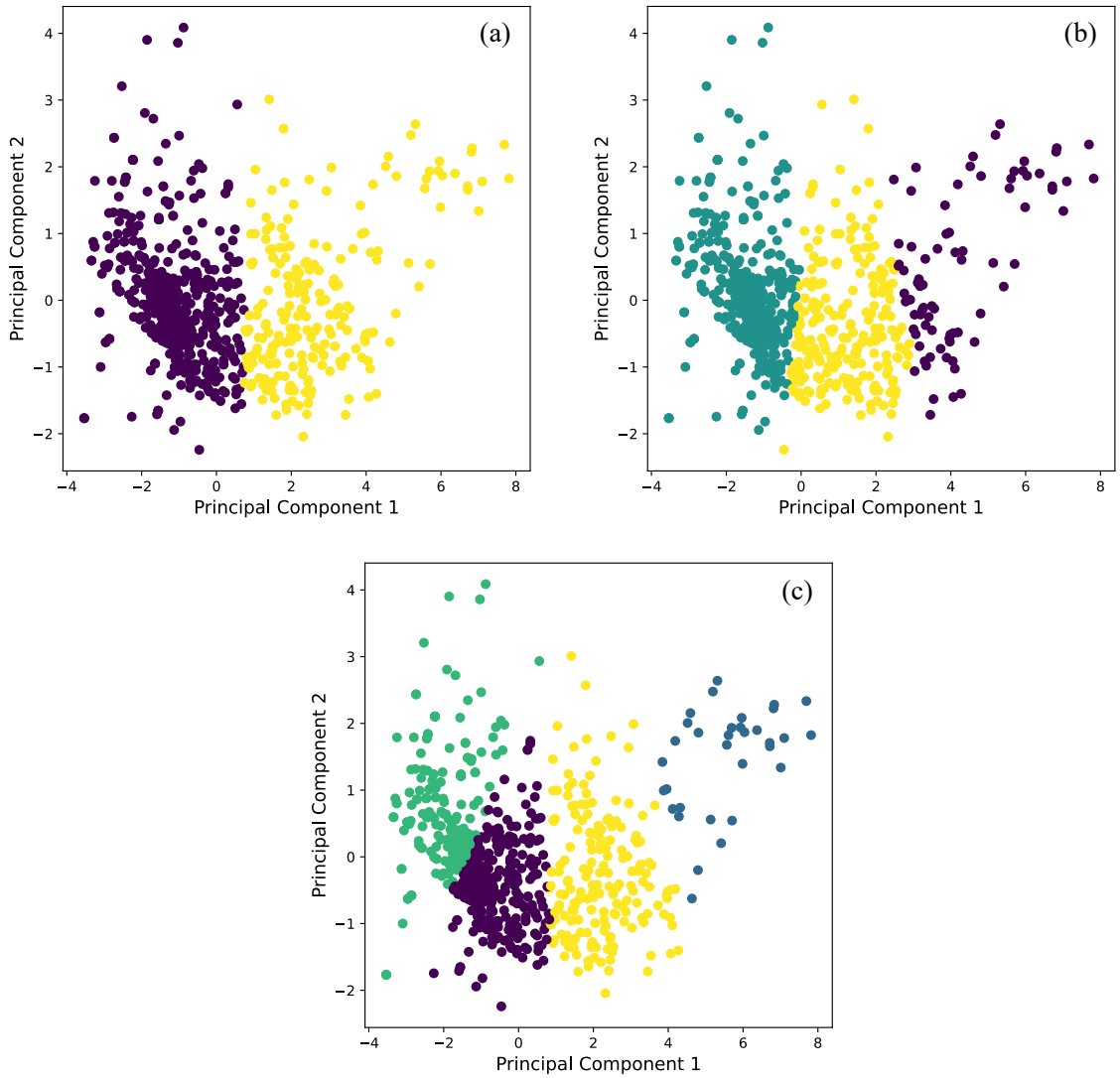

Figure 39. Graph. Cluster analysis of test results based on: (a) two clusters, (b) three clusters, and (c) four clusters.

CHAPTER 3: SUMMARY AND CONCLUSIONS

This study consisted of two major parts. The first part was a literature and specification review aimed at uncovering the current state of the art in terms of multiple stress creep recovery (MSCR) test development and implementation. The second part aimed to develop a robust predictive model for asphalt binder elastic recovery (ER) using the MSCR test results based on a dataset provided by IDOT, so that the potential effectiveness of MSCR as an alternative to asphalt binder ER could be explored. The approach involved exploring ensemble learning methods, specifically tree-based bagging (e.g., random forest and extra trees) and boosting (e.g., AdaBoost, GBDT, XGBoost, LightGBM, and CatBoost). In this context, a hyperparameter tuning process was undertaken to determine the optimal configuration for each model. Additionally, SHAP analysis was employed to assess the feature importance and model interpretation. Moreover, clustering techniques were applied to identify patterns within the dataset. To this end, the dataset, which encompassed six distinct performance grades and a broad spectrum of polymer contents, underwent dimensionality reduction with PCA. Next, the optimal number of clusters was determined using the Silhouette score (2 clusters) and elbow method (2–4 clusters), and the K-means algorithm was applied. The findings of this study are summarized as follows:

- A majority of states in the United States have implemented some form of the MSCR test. In general, states use the same criteria without consideration for low temperature PG or materials specific to the state. Among these states, there is a split between those that have implemented MSCR based on AASHTO M 332 and those that have only used MSCR as a "PG Plus" test. In addition, some require MSCR for all asphalt binders and some only for specific grades, mostly modified asphalt binders.

- The MSCR test has substantial advantages over the current Superpave rutting parameter—namely its assessment of polymer-modified asphalt's ability to recover, even after large deformation, rather than remaining only within the linear viscoelastic range. However, the test in its current form still has shortcomings, including stress levels that remain too low and a lack of sufficient time to fully recover the recoverable portion of the creep response. However, any changes to the test protocol or analysis procedures must be considered in the context of ease of adaptation using current rheometers and testing time.

- The performance of the developed models demonstrated the effectiveness of ensemble learning methods in capturing complex relationships within the data. Extra trees stood out as the top model among the bagging algorithms, while extreme gradient boosting (XGBoost) emerged as the most accurate among the boosting models. Both models exhibited superior predictive accuracy, with extra trees achieving an R^2 value of 0.852 and an RMSE of 2.39 on the testing data. The corresponding values for XGBoost were 0.842 and 2.45, respectively. These findings underscore the potential of ensemble learning for precise predictions of asphalt binder ER from MSCR test results. Remarkably, with the use of these ML methods, the MSCR test can predict the results of the traditional ER test more effectively than the ER-DSR test previously proposed in the literature, despite its different testing temperature range, with MSCR being concentrated on a high temperature (64℃).

- Shapley values analysis provided insights into the feature importance and verified the trustworthiness of the superior ensemble models. Recovery at both stress levels of 0.1 and 3.2 kPa were identified as the top influential features in both extra trees and XGBoost models. Meanwhile, J_{nrdiff} had very little significance, despite its widespread use as a specification parameter, which was consistent with findings in the literature.

- The results of the clustering analysis revealed challenges in distinguishing distinct patterns within the clustered asphalt binders, indicating potential for modifying MSCR analysis as well as further test development to enhance the differentiation of asphalt binders.

- There are several mixture and field validation studies for MSCR in the existing literature, but none focused on Illinois-specific materials, especially considering ongoing research on modifying polymer-modified asphalt binder with softeners. In addition, many tests have focused on validating MSCR using outdated tests such as Marshall flow. A comprehensive rutting study involving validation using the Hamburg wheel-tracking device with Illinois materials is warranted before implementation of an Illinois-specific MSCR specification can be reliably performed when it comes to improving asphalt binder quality.

- However, the implementation of MSCR as a surrogate for ER for IDOT screening of asphalt binder, in terms of determining if an asphalt binder is modified, is appropriate at this time and should be considered due to overall ease of testing. IDOT can also use the predictive algorithm developed in this research study, which will be supplied in the form of a simple web tool, to predict the ER of an asphalt binder based on MSCR, high PG, and polymer content from FTIR.

- It is believed that implementing MSCR in place of ER would not result in a decline in binder quality, as nearly all binders that pass the AASHTO M332 requirement to be considered modified were modified. The only exception is a few asphalt binders that were modified with ground tire rubber (GTR), but not SBS as is conventional in Illinois. However, GTR-modified asphalt binder may have a performance similar to certain SBS-modified asphalt binders, which has not been studied in Illinois so far. On the other hand, MSCR did erroneously categorize some modified asphalt binders as unmodified. For many of these, no ER result was available, while many also passed ER. It is suggested that IDOT run ER as a backup test for an asphalt binder that should be modified but fails MSCR, to be sure before failing it. Furthermore, IDOT can use our developed machine learning tool, which only considered asphalt binders for which the ER result was available, to predict. Therefore, MSCR can partially replace ER, at least as a PG Plus test. This will lead to substantial time savings for IDOT in terms of reducing the dependance on ER, which is a very time-consuming test compared to using the DSR. However, ER has not been analyzed as a tool for examining performance. Therefore, it cannot be stated, as a result of this study, that there is any conclusive way to determine if ER or MSCR is a valuable tool in terms of actual asphalt binder performance, rather than simply detecting the presence of polymer. A comprehensive study with mixture validation that explores the concept of grade bumping would result in a better understanding of the ability of MSCR to predict not only the presence of SBS, which has been established herein, but also the quality of modification and the ability of the test to predict asphalt binder rutting potential,

irrespective of modification state. It is also suggested that the current timing would be very appropriate for such a study, given IDOT is currently evaluating the use of softeners in conjunction with SBS polymer in asphalt binder.

In conclusion, this report highlights the efficacy of ensemble learning methods, particularly extra trees and XGBoost, in accurately predicting asphalt binder elastic recovery from MSCR test results. These models significantly contribute to the understanding of asphalt binder behavior and hold practical implications for states still utilizing ER as a PG Plus test, demonstrating the advantages for them to adopt MSCR into their specifications. Future work will focus on examining relationships between MSCR parameters and mixture performance to determine the best way to implement this test for Illinois, including traffic thresholds.

REFERENCES

AASHTO R 92. (2018). *Standard practice for evaluating the elastic behavior of asphalt binders using multiple stress creep recovery (MSCR) test*. American Association of State Highways and Transportation Officials.

AASHTO T 315. (2022). *Standard method of test for determining the rheological properties of asphalt binder using a dynamic shear rheometer (DSR)*. American Association of State Highways and Transportation Officials.

AASHTO M 320. (2023). *Standard specification for performance-graded asphalt binder*. American Association of State Highways and Transportation Officials.

AASHTO M 332. (2023). *Standard specification for performance-graded asphalt binder using multiple stress creep recovery (MSCR) test*. American Association of State Highways and Transportation Officials.

ASTM C670. (2015). *Standard practice for preparing precision and bias statements for test methods for construction materials*. American Society for Testing and Materials.

ASTM D5801. (2017). *Standard test method for toughness and tenacity of asphalt materials*. American Society for Testing and Materials.

ASTM D6084 (2021). *Standard test method for elastic recovery of asphalt materials by Ductilometer*. American Society for Testing and Materials.

ASTM D8239. (2021). *Standard specification for performance-graded asphalt binder using the multiple stress creep and recovery (MSCR) test*. American Society for Testing and Materials.

Anderson, D. A., & Kennedy, T. W. (1993). Development of SHRP binder specification (with discussion). *Journal of the Association of Asphalt Paving Technologists*, 62.

Asadi, B., & Hajj, R. (2024). Prediction of asphalt binder elastic recovery using tree-based ensemble bagging and boosting models. *Construction and Building Materials*, *410*, 134154.

Bahia, H. U., Hanson, D. I., Zeng, M., Zhai, H., Khatri, M. A., & Anderson, R. M. (2001). *Characterization of modified asphalt binders in superpave mix design* (NCHRP Report 459). National Cooperative Highway Research Program.

de Barros, L. M., do Nascimento, L. A. H., & Aragão, F. T. S. (2022). Effects of binder and aggregate properties on the permanent deformation of asphalt mixtures by means of uniaxial and triaxial tests. *Construction and Building Materials*, 332: 127346. https://doi.org/10.1016/j.conbuildmat.2022.127346

Behnood, A., & Olek, J. (2017). Stress-dependent behavior and rutting resistance of modified asphalt binders: An MSCR approach. *Construction and Building Materials*, 157, 635–646. https://doi.org/10.1016/j.conbuildmat.2017.09.138

Behnood, A., Shah, A., McDaniel, R. S., Beeson, M., & Olek, J. (2016). High-temperature properties of asphalt binders: Comparison of multiple stress creep recovery and performance grading systems. *Transportation Research Record*, *2574*(1), 131–143. https://doi.org/10.3141/2574-15

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24, 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proc. 22nd acm sigkdd Int. Conf. Knowl. Discov. data Min.*, 785–794.

Clopotel, C. S., & Bahia, H. U. (2012). Importance of elastic recovery in the DSR for binders and mastics. *Engineering Journal*, *16*(4), 99–106. https://doi.org/10.4186/ej.2012.16.4.99

D'Angelo, J. A. (2009). The relationship of the MSCR test to rutting. *Road Materials and Pavement Design*, 10 (sup1): 61–80. https://doi.org/10.1080/14680629.2009.9690236

D'Angelo, J. (2010). New high-temperature binder specification using multistress creep and recovery. *Transportation Research Circular n. E-C147*, 1–13.

D'Angelo, J., Kluttz, R., Dongre, R. N., Stephens, K., & Zanzotto, L. (2007). Revision of the Superpave high temperature binder specification: The multiple stress creep recovery test (with discussion). *Journal of the Association of Asphalt Paving Technologists*, 76.

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: Gradient boosting with categorical features support. *arXiv Prepr. arXiv1810.11363*.

Dreessen, S., Planche, J. P., & Gardel, V. (2009). A new performance related test method for rutting prediction: MSCR. *Advanced Testing and Characterization of Bitumen Materials*, *1*, 971–980.

Drucker, H. (1997). Improving regressors using boosting techniques. *Proceedings of the Fourteenth International Conference on Machine Learning*, 107–115.

Feurer, M., & Hutter, F. (2019). Hyperparameter optimization. In F. Hutter, L. Kotthoff, & J. Vanschoren (Eds.), *Automated Machine Learning: Methods, Systems, Challenges* (pg. 3–33). Springer.

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, *14*(5), 771–780.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232.

Gaspar, M. de S., Nogueira, B. G., Vasconcelos, K., Leite, L. F. M., & Bernucci, L. L. B. (2019). Effect of different creep and recovery times on the MSCR test for highly modified asphalt binder. *Journal of Testing and Evaluation*, *49*(1).

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63, 3–42.

Golalipour, A., Bahia, H. U., & Tabatabaee, H. A. (2017). Critical considerations toward better implementation of the multiple stress creep and recovery test. *Journal of Materials in Civil Engineering*, *29*(5), 4016295. https://doi.org/10.1061/(ASCE)MT.1943-5533.0001803

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, 35, 507–520

Gundla, A., Salim, R., Underwood, S. B., & Kaloush, K. E. (2020). Implementation of the AASHTO M 332 specification: A case study. *Transportation Research Record*, *2674*(9), 959–971.

Hajj, R., Filonzi, A., & Bhasin, A. (2019). *Improving the performance-graded asphalt binder specification* (Report No. FHWA/TX-18/0-6925-1). University of Texas at Austin. Center for Transportation Research.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

Ho, T. K. (1995). Random decision forests. *Proc. 3rd Int. Conf. Doc. Anal. Recognit.*, 278–282.

Inocente Domingos, M. D., & Faxina, A. L. (2022). The importance of lower truck speeds (longer creep times) on the rutting responses of polymer-modified asphalt binders. *International Journal of Pavement Research and Technology*, 1–18. https://doi.org/10.1007/s42947-022-00242-2

Hossain, Z., Ghosh, D., Zaman, M., & Hobson, K. (2016). Use of the multiple stress creep recovery (MSCR) test method to characterize polymer-modified asphalt binders. *Journal of Testing and Evaluation*, *44*(1), 507-520.

Jafari, M., Babazadeh, A., & Shahri, M. (2022). The role of stress sensitivity of modified binders with the same linear viscoelastic properties in evaluating rutting resistance of asphalt mixtures. *International Journal of Pavement Engineering*, 1–18. https://doi.org/10.1080/10298436.2022.2027417

Kaloush, K. E., Underwood, B. S., Salim, R., & Gundla, A. (2019). *Evaluation of MSCR testing for adoption in ADOT asphalt binder specifications* (Report No. FHWA-AZ-19-742). Arizona Department of Transportation.

Kataware, A. V., & Singh, D. (2015). Rheological performance of asphalt binders under different creep and recovery periods in MSCR test. *International Journal of Pavement Research and Technology*, *8*(6), 410.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.*, 30.

Liu, H., Zeiada, W., Al-Khateeb, G. G., Ezzet, H., Shanableh, A., & Samarai, M. (2021a). Analysis of MSCR test results for asphalt binders with improved accuracy. *Materials and Structures*, 54, 1–14. https://doi.org/10.1617/s11527-021-01691-0

Liu, H., Zeiada, W., Al-Khateeb, G. G., Shanableh, A., & Samarai, M. (2021b). Use of the multiple stress creep recovery (MSCR) test to characterize the rutting potential of asphalt binders: A literature review. *Construction and Building Materials*, 269, 121320. https://doi.org/10.1016/j.conbuildmat.2020.121320

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.*, 30.

Masad, E. A., Huang, C.-W., DAngelo, J., & Little, D. N. (2009). Characterization of asphalt binder resistance to permanent deformation based on nonlinear viscoelastic analysis of multiple stress creep recovery (MSCR) test. *Journal of the Association of Asphalt Paving Technologists*, 78.

Moraes, R., Swiertz, D., & Bahia, H. (2017). Comparison of new test methods and new specifications for rutting resistance and elasticity of modified binders. *Can. Tech. Asph. Assoc.*

Morshed, M. M. T., Hossain, Z., Chen, D.-H., & Baumgardner, G. (2020). Exploration of alternatives of elastic recovery and conventional fatigue tests of modified binders. *International Journal of Pavement Research and Technology*, 13, 630–636. https://doi.org/10.1007/s42947-020-6009-2

Nanjundan, S., Sankaran, S., Arjun, C. R., & Anand, G. P. (2019). Identifying the number of clusters for K-Means: A hypersphere density based approach. *arXiv Prepr. arXiv1912.00643*.

Olivier, A., Shields, M. D., & Graham-Brady, L. (2021). Bayesian neural networks for uncertainty quantification in data-driven materials modeling. *Computer Methods in Applied Mechanics and Engineering*, 386, 114079.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Adv. Neural Inf. Process. Syst.*, 31.

Radhakrishnan, V., Sri, M. R., & Reddy, K. S. (2018). Evaluation of asphalt binder rutting parameters. *Construction and Building Materials*, *173*, 298–307. https://doi.org/10.1016/j.conbuildmat.2018.04.058

Rajan, B., Suchismita, A., & Singh, D. (2023). Rutting resistance evaluation of highly polymer-modified asphalt binder and mixes using different performance parameters. *Journal of Materials in Civil Engineering*, *35*(8), 04023266. https://doi.org/10.1061/JMCEE7.MTENG-15194

Rao, H., Shi, X., Rodrigue, A. K., Feng, J., Xia, Y., Elhoseny, M., Yuan, X., & Gu, L. (2019). Feature selection based on artificial bee colony and gradient boosting decision tree. *Applied Soft Computing*, 74, 634–642. https://doi.org/10.1016/j.asoc.2018.10.036

Salim, R., Gundla, A., Underwood, B. S., & Kaloush, K. E. (2019). Effect of MSCR percent recovery on performance of polymer modified asphalt mixtures. *Transportation Research Record*, *2673*(5), 308–319. https://doi.org/10.1177/0361198119841283

Singh, D., & Sawant, D. (2016). Understanding effects of RAP on rheological performance and chemical composition of SBS modified binder using series of laboratory tests. *International Journal of Pavement Research and Technology*, *9*(3), 178–189.

Singh, D., Sawant, D., & Xiao, F. (2017). High and intermediate temperature performance evaluation of crumb rubber modified binders with RAP. *Transportation Geotechnics*, 10, 13–21.

Schapire, R. E. (1999, July). A brief introduction to boosting. *Ijcai,* 99(999), 1401–1406.

Stempihar, J., Gundla, A., & Underwood, B. S. (2018). Interpreting stress sensitivity in the multiple stress creep and recovery test. *Journal of Materials in Civil Engineering*, *30*(2), 4017283.

Tabatabaee, N., & Tabatabaee, H. A. (2010). Multiple stress creep and recovery and time sweep fatigue tests: Crumb rubber modified binder and mixture performance. *Transportation Research Record*, *2180*(1), 67–74.

Walubita, L. F., Ling, M., Pianeta, L. M. R., Fuentes, L., Komba, J. J., & Mabrouk, G. M. (2022). Correlating the asphalt-binder MSCR test results to the HMA HWTT and field rutting performance. *Journal of Transportation Engineering, Part B: Pavements*, *148*(3), 04022047.

Wasage, T. L. J., Stastna, J., & Zanzotto, L. (2011). Rheological analysis of multi-stress creep recovery (MSCR) test. *International Journal of Pavement Engineering*, *12*(6), 561–568.

White, G. (2017). Grading highly modified binders by multiple stress creep recovery. *Road Materials and Pavement Design*, *18*(6), 1322–1337. https://doi.org/10.1080/14680629.2016.1212730

Zhang, J., Walubita, L. F., Faruk, A. N. M., Karki, P., & Simate, G. S. (2015). Use of the MSCR test to characterize the asphalt binder properties relative to HMA rutting performance—A laboratory study. *Construction and Building Materials*, *94*, 218–227. https://doi.org/10.1016/j.conbuildmat.2015.06.044

Zoorob, S. E., Castro-Gomes, J. P., Oliveira, L. A. P., & O'Connell, J. (2012). Investigating the multiple stress creep recovery bitumen characterisation test. *Construction and Building Materials*, *30*, 734–745. https://doi.org/10.1016/j.conbuildmat.2011.12.060

Zhou, Z., Gu, X., Dong, Q., Ni, F., & Jiang, Y. (2019). Rutting and fatigue cracking performance of SBS-RAP blended binders with a rejuvenator. *Construction and Building Materials*, *203*, 294–303.

Zhou, F., Li, H., Chen, P., & Scullion, T. (2014). *Laboratory evaluation of asphalt binder rutting, fracture, and adhesion tests* (No. FHWA/TX-14/0-6674-1). Texas Department of Transportation Research and Technology Implementation Office.


Illinois Department
of Transportation


ILLINOIS