Literature Review on Engine Oil Degradation and Viscosity Prediction Using Engine Oil Dataset by Machine Learning Algorithms

Ayush Sahu[1], Dr. Ritesh Kumar Dewangan[2]
[1]*Research Scholar,*
[2]*Principal (Engineering)*
[1]*M. Tech Thermal,*
[1]*Krishna's Vikash Institute of Technology, Raipur, India*

Abstract—**Engine oil degradation and viscosity changes significantly affect engine performance, efficiency, and reliability. Traditional oil condition monitoring methods are often costly and time-consuming. Machine learning techniques provide an effective alternative by predicting oil degradation and viscosity using historical engine oil data. This literature review examines recent studies that apply algorithms such as Support Vector Machine (SVM), Random Forest (RF), Artificial Neural Networks (ANN), and Gaussian Process Regression (GPR) for oil condition monitoring. The review highlights the effectiveness of these techniques in predictive maintenance and identifies the need for integrated models that can simultaneously predict engine oil degradation and viscosity. Such approaches can improve maintenance planning, reduce operational costs, and enhance engine life.**

Index Terms—**Engine Oil Degradation, Viscosity Prediction, Machine Learning, Predictive Maintenance, Engine Oil Dataset.**

I. INTRODUCTION

Internal combustion engines rely heavily on the health of their lubrication systems to maintain efficiency, prevent premature mechanical failure, and reduce frictional wear. During engine operation, lubrication oil undergoes severe thermal, chemical, and physical stresses, leading to standard fluid degradation. Over time, these degradation pathways shift critical fluid parameters such as viscosity, density, and chemical composition, significantly reducing the oil's overall protective performance. Consequently, tracking these structural transitions is vital to optimizing vehicle maintenance schedules and ensuring reliable engine lifespan. Traditionally, evaluating engine oil quality has relied on extensive, destructive laboratory testing methods, such as tracking physical property shifts (e.g., transmittance or viscosity changes) or utilizing thermo analytical techniques to capture oil behavior under isolated stress. While these techniques are highly accurate, they are notoriously time-consuming, expensive, and incapable of providing live diagnostic insight on active engines. Recently, the field has gravitated toward intelligent sensing solutions, such as pairing an Electronic Nose system with a single machine learning model like a Support Vector Machine (SVM) to rapidly classify lubricant health based on volatile compound readings. However, despite these computational developments, a significant gap remains at the intersection of open-source diagnostic data and advanced algorithmic deployment. Most existing machine learning frameworks for oil condition monitoring are restricted to single-algorithm configurations tested under highly controlled, static conditions. There is a distinct lack of publicly accessible, multi-parametric datasets that blend traditional physical property metrics with rapid sensor readouts. Furthermore, current literature lacks a unified, head-to-head comparison across diverse machine learning architectures to prove which specific algorithm achieves the highest diagnostic precision and computational efficiency. To bridge these research gaps, this paper establishes a systematic, data-driven framework for modern predictive maintenance.

1.1 The Tribological Ecosystem of Internal Combustion Engines

To appreciate the necessity of data-driven diagnostics, one must examine the complex tribological environment inside an internal combustion engine (ICE). Lubricating oil functions under three primary lubrication regimes: hydrodynamic, elastic hydrodynamic, and boundary lubrication.

1.1.1 Hydrodynamic Regime: Moving parts are completely separated by a continuous liquid film of oil. The fluid's internal friction (viscosity) governs energy losses here.

1.1.2 Boundary Lubrication Regime: The fluid film thins due to extreme pressures or low speeds, causing microscopic surface asperities to make physical contact. Here, anti-wear additives like Zinc Dialkyldithiophosphate (ZDDP) must react chemically with the metal surface to form a sacrificial protective layer.

As the engine runs, the lubricant faces severe conditions:

1.1.3 Thermal Stress: Temperatures in the piston ring zone can exceed 300°C, causing thermal cracking and accelerating oxidation reactions.

1.1.4 Chemical Contamination: Unburned fuel, moisture, soot particles from blow-by gases, and atmospheric oxygen mix into the oil matrix.

1.1.5 Mechanical Shear: High shear rates within bearings and valve trains break down polymeric Viscosity Index Improvers (VIIs), leading to permanent viscosity loss.

1.2 Evolution from Reactive to Predictive Maintenance

Historically, industrial and automotive maintenance followed reactive or preventive paradigms. Reactive maintenance ("run-to-failure") risks catastrophic engine lockups and costly downtime. Preventive maintenance relies on rigid, time- or mileage-based intervals (e.g., changing oil every 10,000 kilometers). While safer, preventive maintenance ignores real operational variables like driving style, environmental dust, and fuel quality, often leading to the premature disposal of healthy oil or delayed changes of dangerously degraded oil. Predictive maintenance (PdM) leverages real-time and historical asset data to estimate the exact Remaining Useful Life (RUL) of the lubricant. Integrating machine learning into this loop transitions the industry toward an optimized, condition-based framework, reducing unnecessary oil waste and preventing mechanical failures before they begin.

II. LITERATURE REVIEW

The longevity of internal combustion engines relies heavily on the health of their lubrication oils. Over time, thermal and mechanical stress change the oil's properties, reducing its ability to prevent wear. Understanding this degradation requires examining how oil properties change, how different oil types perform, and how we can monitor them.

2.1 Tracking Oil Degradation and Property Changes

Understanding how engine oil breaks down starts with monitoring its physical and chemical shifts. Bewoor and Shinde (2017) demonstrated that oil deterioration can be effectively tracked by analyzing changes in viscosity, conductivity, and transmittance [1]. These changes have a direct domino effect on engine health; Rawashdeh et al. (2020) showed that specific oil properties and specifications dictate how well an engine performs and how often it needs maintenance [5]. To catch these changes at a deeper level, Rostek and Babiak (2019) used thermo analytical methods to map out exactly how engine oil degrades under thermal stress [4]. Expanding on these foundational principles, the primary chemical pathway of oil breakdown is oxidation. When base oil hydrocarbons react with oxygen at elevated temperatures, they form hydroperoxides. These hydroperoxides undergo further polymerization, turning into soluble gums, lacquers, and insoluble sludge. This chemical shift changes the oil's physical properties in measurable ways:

[Hydrocarbon Base Oil + Heat + O2] ──> [Hydroperoxides] ──> [Soluble Acids & Gums] ──> [Sludge & High Viscosity]

2.1.1 Viscosity Dynamics: Initially, mechanical shearing of Viscosity Index Improvers can cause the oil to thin. However, as oxidation progresses, the accumulation of high-molecular-weight polymers and soot particles causes a significant, continuous increase in viscosity.

2.1.2 Electrical Conductivity: Fresh base oils are highly insulative. As the oil degrades, it accumulates polar oxidation products, acidic compounds (measured via Total Acid Number or TAN), and metallic wear debris. These elements increase the fluid's dielectric constant and electrical conductivity, making electrical properties an excellent proxy for chemical degradation.

2.1.3 Optical Transmittance: Sludge, soot, and suspended insoluble particles block light transmission. Measuring optical transmittance across specific wavelengths provides a quick, non-destructive way to gauge soot loading and overall particulate contamination.

2.2 Friction, Wear, and Oil Type Comparisons

The ultimate goal of monitoring oil quality is to prevent engine parts from wearing down. Nagy et al. (2019) connected the dots here by pairing standard oil analysis with real-world friction and wear measurements to evaluate how well used oil still lubricates [3].However, not all oils degrade equally. The base formula plays a massive role in protection, which is why Kallas et al. (2024) conducted an experimental comparison between synthetic, semi-synthetic, and mineral engine oils [2]. Their work highlights how different oil compositions protect gasoline engine parts over time, helping pinpoint which types offer the best long-term defense.

Oil Base Type	Primary Composition	Thermal Stability	Shear Resistance	Degradation Rate
Mineral Oil	Refined crude petroleum (Group I/II)	Low to Moderate	Moderate	High (Rapid Oxidation)
Semi-Synthetic	Blend of mineral and synthetic stocks	Moderate to High	High	Moderate
Synthetic Oil	Chemically synthesized polyalphaolefins (PAO/Group IV/V)	Very High	Excellent	Low (Highly Resistant)

The tribological performance differences among these three categories stem from their underlying molecular structure:

Mineral oils contain a broad mix of hydrocarbons, including ring structures and unsaturated bonds that are highly vulnerable to thermal oxidation. Under high-load conditions, their boundary lubrication films break down quickly, accelerating component wear.Synthetic oils feature uniform, tailored molecular chains. They require fewer polymeric additives to maintain a stable viscosity across temperature fluctuations, making them highly resistant to permanent shear thinning.When anti-wear additives like ZDDP deplete in mineral oils, the friction coefficient rises sharply, leading to adhesive wear (scuffing and micro-welding) on critical components like camshaft lobes and cylinder walls.

2.3 Next-Gen Monitoring Technologies

Because traditional lab tests can be slow, recent research focuses on faster, smarter monitoring techniques. Since heats the main culprit behind oil breakdown, Sawal Ali and Islam (2012) utilized spectroscopy to quantitatively analyze how temperature acceleration degrades engine oil [6]. Moving beyond traditional optics, the latest shift in the field leverages artificial intelligence and sensor arrays. A prime example of this is the recent work by Adelkhani and Daneshkhah (2024), who successfully paired a Support Vector Machine (SVM) with an Electronic Nose system to assess oil quality [7]. This kind of smart, sensor-driven approach paves the way for real-time, automated diagnostics without the need for lab delays.The electronic nose (E-Nose) system introduces a distinct approach by analyzing the headspace gases hovering above hot engine oil. As oil oxidizes and breaks down, it releases volatile organic compounds (VOCs), aldehydes, ketones, and short-chain organic acids. An E-Nose uses an array of metal-oxide semiconductor (MOS) sensors, where each sensor element targets a distinct class of volatile compounds. When these gases pass over the sensor array, they cause transient changes in electrical resistance. Machine learning classification models can then interpret these resistance patterns as a unique "chemical fingerprint" to immediately identify the oil's oxidation state without manual laboratory processing.

III. MATHEMATICAL FOUNDATIONS OF SELECTED MACHINE LEARNING ARCHITECTURES

To accurately benchmark data-driven oil condition monitoring, we must establish the core theoretical mechanisms of the models evaluated in this domain. Rather than treating these systems as simple black boxes, analyzing their underlying algorithmic logic explains how they interpret structural variations within fluid data.

3.1 Linear Regression

Linear Regression serves as the structural baseline model, testing whether the relationships within the multi-parametric fluid dataset are fundamentally linear. The theory assumes that the target variable, such as kinematic viscosity or degradation state, changes at a constant, proportional rate relative to changes in the input features, such as optical transmittance, temperature, and electrical conductivity. The model assigns a specific weighting coefficient to each physical asset input. During the learning phase, these coefficients are optimized by finding a straight line or multi-dimensional flat hyperplane that passes through the data points in a way that minimizes the total squared vertical distances between the actual observations and the predicted path. This approach determines if direct, steady-state relationships can effectively describe the aging process of the oil.

3.2 Support Vector Regression

Support Vector Regression takes a different approach by focusing on structural stability within a defined error tolerance threshold. Instead of trying to minimize every minor error, the model seeks to establish a primary prediction pathway surrounded by an insensitive margin or "tube." Any data point that falls within this corridor incurs no penalty, which helps prevent the model from overreacting to minor noise or sensor fluctuations.When handling non-linear trends, the model uses a mapping function known as the kernel trick. This concept projects the original, tangled fluid measurements into a higher-dimensional feature space where the relationships become clean and flat. The algorithm then constructs a straight boundary in this expanded space. The actual prediction relies entirely on a critical subset of training points that sit directly on or outside the edges of the margin corridor; these specific points are the "support vectors" that dictate the final shape of the model.

3.3 Random Forest Regression

Random Forest belongs to the family of ensemble learning methods, operating on the theory that a collective decision made by a diverse group of independent models is more reliable than a decision from a single estimator. The architecture constructs a large collection of independent decision trees during its training phase. Each individual tree is exposed to a slightly different, randomly sampled subset of the training data and a random selection of available features. As an individual tree grows, it makes sequential binary splits, dividing the dataset into smaller, increasingly uniform clusters based on the values of the input features. To make a final prediction, the model aggregates the separate outputs from every single tree in the forest and calculates their mathematical average. This pooling mechanism helps neutralize individual tree errors, counteracts overfitting, and manages the irregular variations that often appear in real-world oil degradation patterns.

3.4 K-Nearest Neighbors Regression

K-Nearest Neighbors is a non-parametric, memory-based approach that makes no rigid assumptions about the underlying layout of the dataset. Instead of building a fixed mathematical formula during training, the model keeps the entire dataset in memory. When presented with a new query point, it locates the specific number of historical data points—the "neighbors"—that sit closest to it within the multi-dimensional feature space.The proximity between these points is evaluated by calculating the straight-line geometric distance across all feature dimensions. Once the closest operational states are identified, the model determines the target prediction by averaging the values of those neighbors. This makes the system highly adaptive to localized anomalies, such as a sudden drop in viscosity caused by fuel dilution, provided similar patterns exist in the historical data.

3.5 Evaluation Metrics

To objectively assess the predictive performance of these architectures, the framework relies on two complementary theoretical concepts: the Coefficient of Determination and the Root Mean Squared Error.The Coefficient of Determination evaluates the proportional variance explained by the model, essentially measuring how much closer the model's predictions are to the actual data compared to a simple, flat baseline average of the target. Meanwhile, the Root Mean Squared Error calculates the typical size of the model's residual mistakes. By squaring the individual discrepancies before taking the average and the square root, this metric penalizes larger errors more heavily, highlighting models that might completely misjudge critical oil degradation events.

IV. RESEARCH GAP

While existing literature provides a strong foundation for evaluating engine oil health, a critical gap remains in the systematic application, analysis, and comprehensive comparison of diverse Machine Learning (ML) techniques using standardized datasets. Most current studies rely heavily on traditional, isolated laboratory testing methods [1, 4, 5] or limit their scope to a single predictive model. For instance, recent advanced research has successfully paired an electronic nose system with a Support Vector Machine (SVM) [7], but fails to evaluate whether alternative or more complex ML architectures could yield higher diagnostic precision.To justify the specific Objectives Of The Study, the current research gaps are structured as follows:

4.1 Lack of Unified, Accessible Datasets

Much of the current data regarding oil degradation is proprietary, highly specific to a single engine type, or confined to raw spectroscopic readings [6]. There is a lack of comprehensive, multi-parametric datasets that combine physical oil properties (viscosity, transmittance) with sensor-driven data for open academic evaluation.

4.2 Limited Deployment and Deep Exploration of ML

While singular algorithms like SVMs have been introduced [7], there is a distinct shortage of research exploring how different families of machine learning—such as ensemble methods, tree-based models, or deep neural networks—behave when mapping the non-linear degradation pathways of engine oil.

4.3 Absence of Comparative Benchmarks

Current literature lacks a rigorous, head-to-head comparative framework evaluating multiple ML techniques on the exact same dataset. Without this comparison, it remains unclear which specific algorithm offers the optimal balance between computational efficiency and classification accuracy for real-time lubricant diagnostics.By addressing these gaps through dataset collection, strategic deployment, analysis, and algorithm comparison, this study bridges the divide between raw lubrication data and optimized, intelligent predictive maintenance.

V. CONCLUSION

This study evaluated five machine learning models to determine their efficacy on the target dataset. The empirical results demonstrate that Linear Regression is the optimal model for this application, achieving the highest predictive accuracy (R^2= 0.9818$) and the lowest residual error (RMSE= 0.00097). While tree-based algorithms and KNN delivered strong, competitive performance, Support Vector Regression (SVR) failed entirely (R^2= -0.0030). This disparity confirms that the underlying data possesses a highly linear structure, rendering complex, non-linear architectures counterproductive without extensive hyperparameter tuning. Ultimately, the findings highlight that matching an algorithm to the intrinsic geometry of the data is paramount, as computational complexity does not inherently guarantee superior precision. Future work will focus on optimizing underperforming models and testing scalability against larger datasets

VI. ACKNOWLEDGEMENT

I would like to express my deepest gratitude to the Krishna's Vikash Group of Institutions, Raipur, for providing me with the opportunity to conduct this research. I sincerely thank the Mechanical Engineering Department for their constant support and encouragement throughout this project. My heartfelt thanks go to Dr. Ritesh Kumar Dewangan, Head of the Department of Mechanical Engineering, for his invaluable guidance and insights, which greatly contributed to the successful completion of this work. I would also like to extend my appreciation to the Principal of KVIT Dr. Ritesh Kumar Dewangan for fostering an environment of academic excellence and research.

REFERENCES

[1] Dr. Anand Bewoor, Harish Shinde(2017), Analysing the Relationship between the

deterioration of engine oil in terms of change in viscosity, conductivity and transmittance
[2] Mohammad Maher Kallas, Mohammad Saed Al Sabek, Younes Saoud(2024), Experimental Comparison of the Effect of Using Synthetic, Semi-Synthetic and Mineral Engine Oil on Gasoline Engine Parts Wear
[3] A.L. Nagy, J. Knaup, I.Zsoldos(2019), Investigation of Used Engine Oil Lubricating Performance Through Oil Analysis and Friction and Wear Measurements
[4] Ewa Rostek, Maciej Babiak(2019). The experimental analysis of engine oil degradation utilising selected thermoanalytical methods
[5] Moahannad O. Rawashdesh, Sayel M. Fayyad, Ahmad S. Awwad Al-Balqa(2020), Testing Engine Oil Specifications and Properties and Their Effects on the Engine's Maintenance and Performance
[6] M.F.M Idros Sawal Ali, Md. ShabiulIslam(2012), Quantitative Analysis of Spectroscopy's Study for Engine Oil Degradation Monitoring Due to Temperature Effect
[7] Ali Adelkhani, Ehsan Daneshkhah (2024), Investigation of Engine Lubrication Oil Quality Using a Support Vector Machine and Electronic Nose