Enhancing Diamond Clarity Prediction Using Ensemble Machine Learning Models

Shaik Rehaan Ashraf[a], Datla Rohith Varma[b], Sarath S[c]
Department of Computer Science and Engineering,
Amrita School of Computing, Amrita Vishwa Vidyapeetham,
Amritapuri, India
{rehaanashraf2005@gmail.com[a], rohithvarmadatla15@gmail.com[b], saraths@am.amrita.edu[c]}

***Abstract*—Diamond clarity must be assessed not only in the lapidary, but also in high tech in which it affects thermal conductivity, optical precision, and electronic performance. Novel mathematical methods based on releasing bulk information gradients and weak linearizations are simple, and do not require manual inspection or imaging techniques. In order to tackle these problems, this work develops an advanced automated technique which uses ensemble machine learning models (namely Random Forest, LightGBM and XGBoost) to confidently and correctly predict diamond clarity. Our method integrates multi modal data sources such as high resolution images and structured gemstone attributes to improve the clarity classification while dramatically decreasing the reliance on human expertise. We also include innovative feature engineering, use SMOTE for class imbalances, and deploy the interpretability metrics to create a robust and scalable solution for industrial and commercial applications. We experimentally evaluate that our ensemble works better than traditional classification techniques on three diverse datasets, setting a new standard in automated diamond clarity assessment. The transformative promise of machine learning towards gemstone evaluation, grating clarity, is thus brought to light in these findings that assert that machine learning can facilitate faster, more precise and also cheaper evaluation of clarity.**

***Index Terms*—Synthetic Minority Over-sampling Technique (SMOTE), Random Forest (RF), Light Gradient Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost), Diamond Clarity, Semiconductors.**

I. Introduction

Diamonds hold value both for their brilliance in jewelry and their essential roles in high-tech industries, such as semiconductors and precision cutting tools. Clarity, determined by inclusions or structural defects, impacts both value and performance. In jewelry, fewer inclusions lead to higher prices, while in industrial applications, microscopic defects hinder conductivity and optical efficiency. Traditionally, clarity grading depends on manual inspection by gemologists, a subjective and inconsistent process. The same diamond may receive different grades from different evaluators, leading to pricing discrepancies and quality control challenges. As the diamond market expands, scalable and automated grading solutions are increasingly essential.

A major challenge in automating clarity prediction is severe class imbalance in datasets. High-clarity diamonds (e.g., IF, VVS1) are rare, while lower grades are more common. This imbalance causes machine learning models to favor frequent grades, reducing recall for rarer categories. An inaccurate pricing model may undervalue flawless diamonds due to insufficient training data, resulting in financial losses. To address this, we apply the Synthetic Minority Oversampling Technique (SMOTE) to balance the dataset, enabling models to differentiate clarity levels more precisely for fairer predictions.

Another challenge arises from pricing outliers, often caused by grading errors or anomalies. An SI1 diamond mislabeled as IF can skew predictions, leading to overestimated clarity for similar stones. To improve reliability, we implement outlier detection while preserving genuine pricing outliers, as exceptional diamonds naturally command higher prices. Finally, we employ ensemble models—Random Forest, LightGBM, and XGBoost—to integrate gemstone attributes and pricing trends for enhanced clarity prediction. This hybrid approach improves valuation accuracy in gemology and strengthens quality control in industrial applications.

TABLE I
Clarity Grades and Their Applicability in the Semiconductor Industry

Clarity Grade	*Applicability*	*Key Considerations*
SI2	Not Recommended	High inclusion density reduces thermal and electrical efficiency.
SI1	Not Recommended	Defects impact conductivity and optical properties.
VS2	Limited	Fewer inclusions but still unsuitable for high-end applications.
VS1	Conditional	Minimal flaws; possible for heat dissipation.
VVS2	Moderate	High clarity; potential use in thermal management.
Type IIa	Most Suitable	Near-perfect structure; ideal for power electronics and quantum computing.
SCD	Highly Suitable	Excellent purity; widely used in RF and quantum devices.
BDD	Optimal	P-type conductivity; used in sensors and energy storage.
PCD	Effective	Suitable for heat spreaders and industrial coatings.

II. Literature Review

Bendinelli et al. [1] introduced the Gemtelligence study, leveraging deep learning models to classify gemstones using

multimodal data. Their approach significantly reduces reliance on costly and destructive laboratory testing by incorporating spectral and elemental analysis, offering a scalable method for gemstone origin determination and treatment detection. However, their study primarily focuses on authenticity verification rather than diamond clarity assessment, which is crucial for valuation. Additionally, the scalability of their approach is constrained by high computational requirements and dataset variability, limiting its applicability in real-world clarity grading systems.

Fitriani et al. [2] investigated machine learning efficiency in price prediction by applying Label Encoding to convert clarity grades into numerical values. Their study compared k-Nearest Neighbors (k-NN) and LASSO regression, using SelectKBest to identify key attributes such as carat, x, y, and z. The results showed that k-NN outperformed LASSO, achieving a lower RMSE (926.06) and a higher R^2 (0.9066). However, the study does not explore the underlying reasons for k-NN's superior performance or analyze how clarity interacts with other attributes in price prediction. Furthermore, external market factors, diamond quality variations, and social influences impact prediction accuracy. To improve real-world deployment, future studies should incorporate advanced hyperparameter tuning and enhanced image preprocessing to address challenges such as blurry certificates and incorrect attribute values.

Chow and Reyes-Aldasoro [3] applied computer vision and machine learning to gemstone classification, utilizing image-based feature extraction to improve accuracy and reduce classification time. Their model outperformed human experts (69.4% vs. 66.9%), but it was primarily designed for gemstone type identification rather than clarity assessment. Given that clarity grading requires fine-grained defect detection of internal inclusions and blemishes, existing models that focus on color and texture analysis lack the necessary capabilities. Although deep learning techniques such as ResNet show potential, a significant research gap remains in automating clarity evaluation using microscopy, advanced image processing, and multimodal learning. Future research should explore CNNs, high-resolution imaging, and spectroscopy to improve clarity grading accuracy, making automation more reliable and scalable in the gemstone industry.

Mihir et al. [4]emphasized the importance of diamond clarity in valuation, graded on an eight-point scale from I1 to IF, alongside the 4Cs—carat, cut, color, and clarity—that define diamond grading standards. Their research demonstrated high precision in price prediction, with Random Forest Regression achieving an R^2 of 97.93%, while the M5P model improved from 98.7% to 99.03% through dimensionality reduction. However, a key research gap remains in systematically comparing regression models for a GUI-based price prediction system using scanned certificates. Additionally, feature correlation issues, blurry certificate images, and incorrect attribute values hinder prediction accuracy. A more scalable, automated approach integrating clarity assessment with price estimation is necessary to enhance practical usability in the diamond industry.

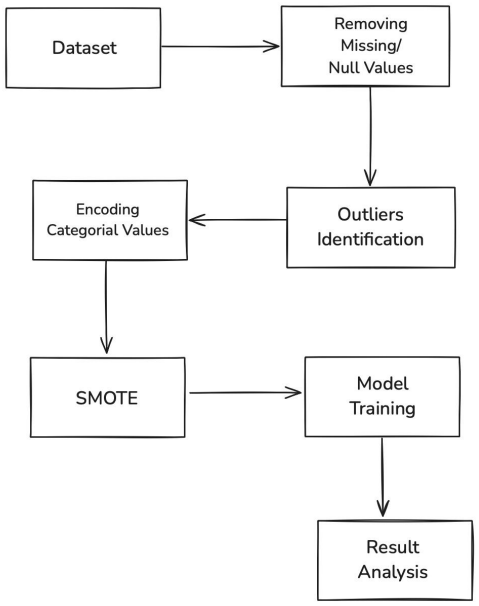

Fig. 1. Workflow of Proposed model

III. Proposed Methodology

In this study, we conduct a comprehensive evaluation of ensemble learning models for diamond clarity classification, utilizing three state-of-the-art algorithms: Random Forest, LightGBM, and XGBoost. To ensure robustness and generalizability, we analyze diverse datasets sourced from publicly available repositories such as Kaggle, PyCaret, and GitHub. These datasets encompass variations in feature distributions and data quality, enhancing the reliability of our findings.

A. Dataset

This study evaluates model performance on three heterogeneous diamond clarity prediction datasets, sourced from publicly accessible repositories to ensure diversity in scale and feature representation:

- Dataset-A (Github): Comprises 53,940 records with 10 attributes, including 8 numerical (e.g., carat, depth, price) and 2 categorical features (cut, clarity). All features are ,complete with no missing values.

	carat	cut	depth	table	price	x	y	z	Clarity	Color
0	0.23	1	61.5	55.0	326	3.95	3.98	2.43	SI2	E
1	0.21	2	59.8	61.0	326	3.89	3.84	2.31	SI1	E
2	0.23	4	56.9	65.0	327	4.05	4.07	2.31	VS1	E
3	0.29	2	62.4	58.0	334	4.20	4.23	2.63	VS2	I
4	0.31	4	63.3	58.0	335	4.34	4.35	2.75	SI2	J

Fig. 2. Dataset-A Overview

- Dataset-B (PyCaret): Contains 6,000 records and 8 attributes, dominated by categorical variables (6 features,

e.g., color, cut) and 2 numerical attributes (carat, price). No missing values are present.

	Carat Weight	Cut	Color	Clarity	Polish	Symmetry	Report	Price
0	1.10	Ideal	H	SI1	VG	EX	GIA	5169
1	0.83	Ideal	H	VS1	ID	ID	AGSL	3470
2	0.85	Ideal	H	SI1	EX	EX	GIA	3183
3	0.91	Ideal	E	SI1	VG	VG	GIA	4370
4	0.83	Ideal	G	SI1	EX	EX	GIA	3171

Fig. 3. Dataset-B Overview

- Dataset-C (Kaggle): Consists of 26,967 records and 11 attributes, with 7 numerical (e.g. x, y, z dimensions) and 3 categorical features (clarity, color, cut). One column exhibits missing values, addressed via median imputation for numerical features and mode-based imputation for categorical features during preprocessing.

	S.No	carat	cut	color	clarity	depth	table	x	y	z	price
0	1	0.30	Ideal	E	SI1	62.1	58.0	4.27	4.29	2.66	499
1	2	0.33	Premium	G	IF	60.8	58.0	4.42	4.46	2.70	984
2	3	0.90	Very Good	E	VVS2	62.2	60.0	6.04	6.12	3.78	6289
3	4	0.42	Ideal	F	VS1	61.6	56.0	4.82	4.80	2.96	1082
4	5	0.31	Ideal	F	VVS1	60.4	59.0	4.35	4.43	2.65	779

Fig. 4. Dataset-C Overview

All data sets adhere to industry standard gemological attributes, including carat, cut, color, clarity, depth, table, price, and dimensional measurements (x, y, z), ensuring a consistent predictive feature space. Dataset heterogeneity, reflected in varying record counts, attribute distributions, and categorical-to-numerical ratios, enables a robust evaluation of model generalizability across different data regimes. The absence of missing values in Dataset-A and Dataset-B, along with systematic handling in Dataset-C, minimizes bias in comparative analysis.

This curated dataset collection facilitates a rigorous evaluation of ensemble methods under various conditions, capturing real-world variations in gemological data quality and feature engineering practices.

B. Feature Description

This study examines a diamond dataset with numerical and categorical attributes influencing quality and valuation. Table II summarizes these features. Numerical attributes—Carat, Depth, Table, Price, X, Y, and Z—capture physical dimensions and cost. Carat determines size, while Depth and Table affect brilliance. X, Y, and Z define length, width, and height, while Price reflects market value. Categorical attributes—Cut, Clarity, Color, Polish, Symmetry, and Report—impact grading and certification. Cut influences brilliance, Clarity assesses inclusions, and Color measures transparency, with higher grades being more colorless. Polish and Symmetry determine finishing precision, affecting appearance. The Report attribute certifies diamonds through gemological institutes, ensuring standardized evaluation. Understanding these attributes helps analyze pricing trends, quality grading, and factors contributing to a diamond's desirability. .

TABLE II
DIAMOND FEATURES AND THEIR RANGES

Feature	***Range***
Carat Weight	0.2 – 5.01
Cut	Categorical (Fair, Good, Very Good, Premium, Ideal)
Depth	43 – 79
Table	43 – 95
Price	326 – 18,823
x (length in mm)	0 – 10.74
y (width in mm)	0 – 58.9
z (depth in mm)	0 – 31.8
Clarity	Categorical (I1, SI2, SI1, VS2, VS1, VVS2, VVS1, IF)
Color	Categorical (D, E, F, G, H, I, J)
Polish	Categorical (Fair, Good, Very Good, Excellent, Ideal)
Symmetry	Categorical (Fair, Good, Very Good, Excellent, Ideal)
Report	Categorical (e.g., GIA, AGSL)

C. Data Preprocessing

The datasets underwent systematic preprocessing to minimize biases and ensure consistency between features. We inspect missing values in all datasets. Data sets A and B contained no null entries, while data set C had missing values in one column. These were addressed using median imputation for numerical features and mode-based substitution for categorical attributes.

To standardize clarity prediction across datasets, textual clarity grades (for example, IF, VVS1, VS2) were assigned to ordinal numerical values reflecting their hierarchical significance in gemological standards, as outlined in Table 1. This transformation preserved the intrinsic quality hierarchy of the clarity grades while enabling numerical compatibility for model training.

D. Outlier Detection and Handling

Outliers were identified using the Interquartile Range (IQR) method, a standard statistical technique that effectively handles skewed distributions, making it well-suited for diamond pricing data. While alternatives like Z-score and DBSCAN clustering were considered, Z-score was found to be highly sensitive to non-normal distributions, and DBSCAN struggled with varying density levels in the dataset. IQR provided a more stable approach by detecting extreme values based on quartile boundaries.

Outliers were detected across multiple attributes, with 3,523 in dataset A, 1,419 in dataset B, and 1,779 in dataset C. Attributes such as carat, table size, and depth contained inconsistencies or measurement errors, leading to their removal. However, outliers in price were retained since high-clarity diamonds naturally command premium values. For example, a

flawless 3-carat diamond may be priced exponentially higher than a lower-clarity counterpart of the same size. Removing these values would distort the model's understanding of pricing trends and lead to inaccurate predictions.

A threshold of 1.5 times the interquartile range was used for detection, ensuring a balance between filtering extreme anomalies and preserving valid variations. By selectively removing outliers while retaining relevant pricing extremes, the model maintains alignment with industry valuation practices. This approach improves predictive accuracy by preventing bias while ensuring realistic price estimations for high-value diamonds.

E. Encoding

The categorical features, including cut, color, and clarity, were encoded using domain-specific strategies to retain their semantic relationships and improve model interpretability. Clarity, being an ordinal variable with a natural hierarchy (I1 $<$ SI2 $<$ SI1 $<$ VS2 $<$ VS1 $<$ VVS2 $<$ VVS1 $<$ IF), was mapped numerically from 0 to 7 to preserve its ranking, ensuring that machine learning models could recognize the progressive nature of clarity improvements. Meanwhile, nominal variables like color and cut, which have no intrinsic order, were encoded using one-hot encoding to prevent the model from assuming an ordinal relationship where none exists. This dual encoding strategy optimizes the performance of ensemble models by allowing them to capture feature interactions effectively. Ordinal encoding enables models like LightGBM and XGBoost to leverage gradient-based prioritization in decision trees, while one-hot encoding ensures that categorical splits remain distinct, preventing misinterpretations due to artificial ordinal relationships. This hybrid approach enhances predictive accuracy and model stability.

F. Class Imbalance

As shown in Figure 5, The diamond clarity datasets (Dataset-A, Dataset-B, and Dataset-C) exhibit significant imbalance, with rare clarity grades (e.g., IF) being underrepresented. This affects model performance, as ensemble models like Random Forest, LightGBM, and XGBoost struggle with minority classes. In Dataset-A, Random Forest achieved 62.26% accuracy but only 58% recall for the minority class. LightGBM and XGBoost had recall values between 52% and 55%, despite overall accuracies of 61%-63%. Additionally, the datasets show no consistent pattern. Dataset-A and Dataset-C share similar distributions, with mid-range clarity levels (SI1, VS2) being dominant, while Dataset-B has an entirely different distribution, with no instances of the IF clarity grade. This variability highlights real-world inconsistencies, making model generalization challenging.

To improve reliability, the Synthetic Minority Oversampling Technique (SMOTE) was applied to generate synthetic samples for underrepresented clarity grades while preserving ordinal relationships critical for gemological accuracy. By balancing the dataset, SMOTE ensures models do not disproportionately favor majority classes, leading to fairer and more accurate predictions. This approach enhances research reliability by ensuring that models trained on one dataset can generalize effectively across varying data distributions. Addressing class imbalance through SMOTE improves predictive fairness, making automated clarity grading systems more trustworthy and applicable to diverse datasets, reducing bias, and strengthening the validity of the research findings.

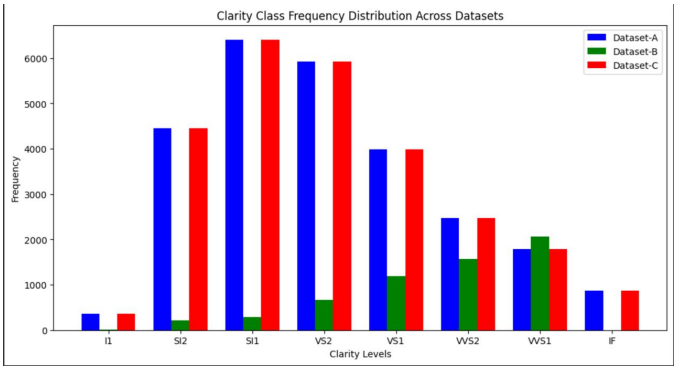

Fig. 5. Clarity Class Frequency Distribution Across Datasets

G. Machine Learning Models

Classifying diamond clarity accurately requires models that can manage class imbalance while maintaining predictive reliability. The chosen ensemble methods—Random Forest, LightGBM, and XGBoost—were selected for their effectiveness in handling imbalanced data, interpretability, and efficiency. Random Forest provides stability, LightGBM enhances speed, and XGBoost improves generalization with regularization. These models collectively balance accuracy, computational efficiency, and the challenge of rare clarity grades.

1) Random Forest (RF): Uses multiple decision trees built through bootstrap sampling and feature randomness. It aggregates predictions via majority voting (classification) or averaging (regression) to reduce variance and overfitting. RF determines feature importance using Gini impurity reduction, highlighting key factors like carat, cut, and depth. The model was set up with `n_estimators` = 150, `min_samples_split` = 5, `max_depth` = 20, and `bootstrap` = True.
2) LightGBM (Light Gradient Boosting Machine): A gradient-boosting framework designed for large datasets. Unlike level-wise growth, LightGBM splits nodes leaf-wise, prioritizing those that maximize loss reduction. This approach improves both efficiency and accuracy, while weighted sampling helps mitigate class imbalance, making it suitable for identifying rare clarity grades. The model parameters were `n_estimators` = 150, `max_depth` = 8, and `learning_rate` = 0.1.
3) XGBoost (Extreme Gradient Boosting): An optimized gradient-boosting algorithm that enhances performance through second-order gradient approximations and L1/L2 regularization, reducing overfitting. XGBoost pri-

oritizes node splits based on gradient-based heuristics, optimizing selection while handling sparse data effectively. The `scale_pos_weight` parameter improves recall for underrepresented clarity grades. The model was configured with `learning_rate` = 0.1, `max_depth` = 9, and `n_estimators` = 200.

RESULTS

For clarity classification, three diamond datasets with varying class distributions and outliers were analyzed. SMOTE enhanced minority class representation, helping models handle clarity differences. Precision, recall, and F1-score were used for evaluation. Precision measures correctly predicted positives, reducing false positives. Recall assesses actual positives identified, minimizing false negatives. F1-score, the harmonic mean of precision and recall, balances both. These metrics ensured a comprehensive assessment of Random Forest, LightGBM, and XGBoost across datasets.

- Table III shows the performance of Random Forest, LightGBM, and XGBoost after SMOTE on Dataset-A, where class frequencies ranged from 355 (I1) to 6408 (SI1). LightGBM and XGBoost excel in high-frequency classes before SMOTE, such as SI1 (6408) and SI2 (4447), achieving high F1 scores due to gradient-based optimization. Random Forest remains stable in lower-frequency classes like I1 (355) and IF (874), leveraging bagging to prevent overfitting. In middle-frequency classes before SMOTE, like VS1 (3991) and VVS2 (2479), all models perform similarly, with LightGBM and XGBoost showing slight advantages. Overall, LightGBM is best for high-frequency classes, XGBoost follows closely, and Random Forest performs better in rare classes, handling SMOTE-generated data effectively.

TABLE III
PERFORMANCE COMPARISON OF DATASET-A

Class	***Random Forest***			***LightGBM***			***XGBoost***		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
0	0.96	1.00	0.98	0.98	0.99	0.99	0.98	1.00	0.99
1	0.80	0.81	0.81	0.83	0.83	0.83	0.83	0.82	0.83
2	0.62	0.68	0.65	0.63	0.70	0.66	0.63	0.70	0.67
3	0.67	0.59	0.63	0.66	0.61	0.64	0.68	0.62	0.65
4	0.67	0.66	0.67	0.67	0.64	0.65	0.67	0.66	0.66
5	0.76	0.74	0.75	0.75	0.75	0.75	0.77	0.76	0.76
6	0.80	0.77	0.79	0.81	0.79	0.80	0.83	0.81	0.82
7	0.88	0.92	0.90	0.90	0.92	0.91	0.91	0.92	0.91

- Table IV presents the performance of Random Forest, LightGBM, and XGBoost after SMOTE was applied to Dataset-B, where initial class distributions were highly imbalanced, with only 4 samples for I1 and 0 for IF. SMOTE significantly improved representation, enabling LightGBM to achieve superior F1 scores in moderately sized classes before SMOTE, such as VS2 (666) and VS1 (1192), reaching 0.83 and 0.86, respectively. XGBoost demonstrated strength in capturing finer clarity distinctions, particularly in VVS2 (1575) and VVS1 (2059), with F1 scores of 0.75 and 0.77. Random Forest, while maintaining stable predictions, had slightly lower F1 scores due to its averaging mechanism, making it less adaptive to complex decision boundaries. All three models performed optimally in well-represented classes, with F1 = 1.00 for Class 0 across all models, while LightGBM and XGBoost showed slight advantages in handling the effects of SMOTE on previously underrepresented classes.

TABLE IV
PERFORMANCE COMPARISON OF DATASET-B

Class	**Random Forest**			**LightGBM**			**XGBoost**		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1	0.95	0.94	0.94	0.94	0.96	0.95	0.96	0.95	0.95
2	0.90	0.87	0.89	0.90	0.89	0.89	0.90	0.90	0.90
3	0.75	0.83	0.78	0.81	0.80	0.80	0.81	0.83	0.82
4	0.63	0.59	0.61	0.63	0.63	0.63	0.67	0.65	0.66
5	0.62	0.60	0.61	0.68	0.64	0.66	0.69	0.69	0.69
6	0.78	0.82	0.80	0.85	0.90	0.87	0.86	0.88	0.87

- Table V presents the performance of Random Forest, LightGBM, and XGBoost after SMOTE was applied to Dataset-C, which had the same class distribution as Dataset-A. XGBoost demonstrated stronger performance in handling complex clarity grades, particularly in VVS2 (2479), where it achieved an F1 score of 0.77, surpassing LightGBM at 0.75. Boosting methods effectively captured subtle variations, with XGBoost maintaining higher F1 scores in middle-frequency classes before SMOTE, such as VS2 (5925) and VS1 (3991). Random Forest remained stable across the dataset, reinforcing its strength in interpretability and mitigating overfitting. In highly frequent classes before SMOTE, such as SI1 (6408), LightGBM and XGBoost achieved the highest F1 scores, leveraging gradient-based optimization. Random Forest performed reliably in lower-frequency classes, balancing variance through bagging. Overall, XGBoost excelled in complex cases, LightGBM in well-represented classes, and Random Forest ensured consistent predictions across clarity levels.

TABLE V
PERFORMANCE COMPARISON OF DATASET-C

Class	***Random Forest***			***LightGBM***			***XGBoost***		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
0	0.96	0.99	0.98	0.98	0.99	0.99	0.97	1.00	0.98
1	0.79	0.82	0.80	0.80	0.83	0.82	0.83	0.82	0.83
2	0.61	0.67	0.64	0.59	0.69	0.64	0.64	0.70	0.66
3	0.66	0.57	0.61	0.64	0.56	0.60	0.66	0.61	0.63
4	0.63	0.63	0.63	0.62	0.55	0.59	0.64	0.62	0.63
5	0.72	0.69	0.70	0.69	0.64	0.66	0.75	0.74	0.74
6	0.75	0.74	0.74	0.70	0.72	0.71	0.80	0.78	0.79
7	0.85	0.89	0.87	0.84	0.91	0.87	0.89	0.92	0.90

FUTURE SCOPE

In the future, we can increase the accuracy of the diamond clarity prediction model by using the following terms:

- Advanced Feature Engineering: The development of improved feature engineering techniques should be a main research goal for the future because it aims to extract better features from raw multimodal data sets that include detailed images and structured stone information. Researchers should pursue the development of complex image processing algorithms as well as the creation of new attributes that comprehensively represent diamond clarity features.
- Hybrid Modeling Approaches: Integrating different machine learning paradigms offers a promising avenue for improving predictive accuracy. Future work could explore hybrid models that combine ensemble methods such as Random Forest, LightGBM, and XGBoost with deep learning techniques to leverage their respective strengths.
- Expanded Gemological Applications: Research related to diamond clarity serves as a foundation for developing methodologies that can be applied to evaluate different types of gemstones. The study proposes additional research into the potential of ensemble learning models with multimodal inputs to support color grading along with cut quality assessment, treatment detection, and origin verification as described in Bendinelli et al. [1].

CONCLUSION

In conclusion, this study demonstrates the effectiveness of ensemble learning methods in automating diamond clarity classification across diverse datasets. By integrating robust data preprocessing steps, including missing value imputation, appropriate categorical feature encoding, and the application of SMOTE to address class imbalance, our approach ensured that the intrinsic complexities of gemological data were preserved. The comparative analysis across three distinct datasets revealed that boosting algorithms, particularly LightGBM and XGBoost, excel in capturing subtle clarity variations in classes with ample samples, while Random Forest maintains stable performance in scenarios with limited representation. These findings underscore the importance of leveraging ensemble techniques to manage non-linear relationships and mitigate overfitting in high-dimensional, heterogeneous datasets. The proposed methodology not only enhances predictive accuracy but also offers practical insights for industrial applications where diamond quality assessment is critical. Future research may explore the integration of advanced feature engineering and hybrid modeling approaches to further refine classification performance and extend applicability to broader gemological evaluations.

REFERENCES

[1] Bendinelli, T., Biggio, L., Nyfeler, D., Ghosh, A., Tollan, P., Kirschmann, M. A., & Fink, O. (2024). GEMTELLIGENCE: Accelerating gemstone classification with deep learning. *Communications Engineering, 3*(1), Article 110. Available at: https://doi.org/10.1038/s44172-024-00252-x.

[2] Fitriani, S. A., Astuti, Y., & Wulandari, I. R. (2022). Least Absolute Shrinkage and Selection Operator (LASSO) and k-Nearest Neighbors (k-NN) Algorithm Analysis Based on Feature Selection for Diamond Price Prediction. In *Proceedings of the 2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)* (pp. 135-139). IEEE. Available at: http://dx.doi.org/10.1109/ISMODE53584.2022.9742936.

[3] Chow, B. H. Y., & Reyes-Aldasoro, C. C. (2022). Automatic Gemstone Classification Using Computer Vision. *Minerals, 12*(1), 60. Available at: https://doi.org/10.3390/min12010060.

[4] Mihir, H., Patel, M. I., Jani, S., & Gajjar, R. (2021). Diamond Price Prediction using Machine Learning. In *Proceedings of the 2021 2nd International Conference on Communication, Computing and Industry 4.0 (C2I4)* (pp. 1-6). IEEE. Available at: http://dx.doi.org/10.1109/C2I454156.2021.9689412.

[5] Amadavadi, K., Rane, R., & Patankar, R. (2024). Diamond Price Prediction using Machine Learning Techniques. In *Proceedings of the 2024 5th International Conference on Computing, Communication, and Cyber-Security (IC4S)* (pp. 1-6). IEEE. Available at: http://dx.doi.org/10.1109/ICOSEC61587.2024.10722317.

[6] Basha, M. S. A., Oveis, P. M., Prabavathi, C., Lakshmi, M. B., & Sucharitha, M. M. (2023). An Efficient Machine Learning Approach: Analysis of Supervised Machine Learning Methods to Forecast the Diamond Price. In *Proceedings of the 2023 International Conference for Advancement in Technology (ICONAT)* (pp. 1-6). IEEE. Available at: https://doi.org/10.1109/ICONAT57137.2023.10080618.

[7] Sarath, S., & Nair, J. J. (2024). Detection and Classification of Respiratory Syndromes in Original and Modified DCGAN Augmented Neonatal Infrared Datasets. *Procedia Computer Science, 233*, 422-431. Available at: https://doi.org/10.1016/j.procs.2024.03.232.

[8] Teki, V. R. N. M., Ragaven, R. A., Manoj, N., V. V., & S. S. (2023). A Comparison of Two Transformers in the Study of Plant Disease Classification. In *Proceedings of the 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE. Available at: https://doi.org/10.1109/ICCCNT56998.2023.10307591.

[9] Avuthu, B., Yenuganti, N., Kasikala, S., Viswanath, A., & Sarath, S. (2022). A Deep Learning Approach for Detection and Analysis of Respiratory Infections in COVID-19 Patients Using RGB and Infrared Images. In *Proceedings of the 2022 Fourteenth International Conference on Contemporary Computing (IC3)* (pp. 1-7). IEEE. Available at: https://doi.org/10.1145/3549206.3549272.

[10] Chaitanya, S., Sarath, S., Malavika, Prasanna, & Karthik. (2020). Human Emotions Recognition from Thermal Images using YOLO Algorithm. In *Proceedings of the 2020 International Conference on Communication and Signal Processing (ICCSP)* (pp. 1139-1142). IEEE. Available at: https://doi.org/10.1109/ICCSP48568.2020.9182148.

[11] M. S. . K. Usha Rani, "Enhancement of Cloud Data Protection using Attribute Based Encryption with Multiple Keys: A Survey", MSEA, vol. 72, no. 1, pp. 1952–1967, Jul. 2023.

[12] Vaishnavi, K. P., Ramadas, M. A., Chanalya, N., Manoj, A., & Nair, J. J. (2021). Deep Learning Approaches for Detection of COVID-19 Using Chest X-Ray Images. In *Proceedings of the 2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT)* (pp. 1-8). IEEE.Available at: http://dx.doi.org/10.1109/ICECCT52121.2021.9616623.

[13] Krishnendu, S., Sarath, S., Niveditha, S., Siva Sai, B. M. (2024). Heart Failure Prediction using Machine Learning Techniques: A Comparative Analysis. *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-6). Available at: http://dx.doi.org/10.1109/ICCCNT61001.2024.10724418