Grenze International Journal of Engineering and Technology, June Issue

Compliance-First Hybrid Framework for Automated Insurance Claim Verification

Sangita Jaybhaye[1], Divesh Lokhande[2], Kumar Kolhe[3], Rahul Kuwar[4], Murali Kulkarni[5] and Omkar Kharate[6]
[1-6]Computer Science & Engineering (AI), Vishwakarma Institute of Technology, Pune, India
Email: sangita.jaybhaye@vit.edu, divesh.lokhande24@vit.edu, kumar.kolhe24@vit.edu, rahul.kuwar24@vit.edu, murali.kulkarni24@vit.edu, omkar.kharate24@vit.edu

***Abstract*— Checking health insurance claims is mandatory but it is time consuming process. If a claim is checked manually, then it will increase the operational cost and take more time for patient settlement. Many of the existing automated systems are based on either strict rule-based checks or data-driven models only, which decrease their capacity to process complicated logic of policy.**
We made a hybrid machine learning system that mixes structured claim information, rule-based validation, and domain-aware attributes of user behavior, document reliability, risk indicators, and policy compliance. After doing dataset cleaning and feature engineering, a gradient boosting then our model categorizes claims into acceptable, not acceptable, or needing more information. The model gained high consistency (around 87%) and scalability over conventional rule-based approaches.

***Keywords*— Insurance Claim Verification· Machine Learning· Gradient Boosting· Domain-Informed Feature Engineering· Risk Assessment· Claim Classification.**

I. INTRODUCTION

Re-checking of health insurance claims is a part of what insurance companies do. They have to follow a lot of rules and be careful when they look at claims. As more and more claims come in insurance companies need to process them without breaking any rules or making decisions that are not clear. Checking claims by hand is a lot of work can be uneven and cannot be done for a number of claims. We need systems that can check claims automatically and follow the rules.
Current automated systems are either rule-based or use machine learning. Rule-based systems have rules but they have trouble with complicated cases and spotting fraud.
Machine learning models are good at predicting what might happen and spotting things that're not normal but they do not always follow the rules and can be hard to use in some situations. To fix this problem we suggest a system that puts following the rules first.
This system has two parts: one that checks the rules and another that looks at the chances of something happening. It starts with a rule engine that applies the policy rules. Claims that meet these rules are then looked at by a kind of classifier called Extreme Gradient Boosting, which looks at things like money ratios what the person claiming has done before if they followed the rules and how good the documents are. If a claim is not clear it gets sent to a human to look at. If it is rejected the person who made the claim gets feedback on why it was rejected and what rules were broken.

Grenze ID: 01.GIJET.12.2.697_26

The main things we are doing in this work are:

- –A Making a system that follows the rules and uses machine learning to make predictions so the rules are not broken.– Using a kind of classifier to look at claims and decide what category they fit into based on things like money ratios and if the person claiming has followed the rules before.
- Making a system that sends claims to humans to look at if it's not sure and gives feedback on what rules were broken and what was important, in making the decision.
- Comparing different machine learning algorithms to see which one works best: Logistic Regression, Decision Tree, Random Forest and Extreme Gradient Boosting.

II. Literature Review

Conventional health insurance claim processing systems use predefined rules to ensure policy compliance and check document completeness. While they meet regulatory requirements and are easy to understand, they do not adjust easily to evolving fraud schemes and complex unusual cases [9, 10]. Several machine learning methods have focused on healthcare and insurance claim analytics to improve flexibility, deriving decision boundaries from structured claim data [2, 4]. These all models will look at claims. And they will put them into groups based on fraud trends that are happened in the past. These models will use methods, and gradient boosting like models will do the best job with dataset which are organized into the tables. Friedman is the man who figured out the idea, behind all of this. [1], and mainly scalable methods like XGBoost can offer increment in the various parameters like generalization

Sangita Jaybhaye, Divesh Lokhande, Kumar Kolhe, Rahul Kuwar, Murali Kulkarni, and Omkar Kharate and computational speed [3]. The performance of prediction is mostly dependent on the feature engineering. Some previous studies have shown that including the claimant behavior statistics, the policy conformity metrics and the historical risk scoring is really essential for the prediction. The feature engineering is very important because it helps to make the prediction better. The claimant behavior statistics the policy conformity metrics and the historical risk scoring are all things to include when you are doing the feature engineering, for the prediction. [4, 5]. Now there are more ways for us to have access to information through the use of documents. One example is where we can access documents like scanned insurance papers using computerized programs that can gather information from documents that are scanned to give us the data we need about the layout of an insurance paper. Therefore, we will be able to get the information from insurance documents using OCR (Optical Character Recognition) pipelines in addition to layout-aware transformer models to access the data [6, 7], However, these systems typically function without regard to enforcement mechanisms. When combined, deterministic rule based systems and probabilistic machine learning offer a balance of trade through enhanced reliability and restricting user defined policies prior to the development of statistical models. [4, 8]. Regardless of such advances, very little research has merged text recognition, manual verification, rule verification, and foreseeable direction into a single pipeline. This gap motivates the proposed hybrid framework. Despite advancements in technology, there is still a considerable amount of migration, research, and minimal studies incorporating text-to-speech identification, virtually no verification method for documentation that utilizes the rules for printing, and very few predictive analytics models. Within that same domain there are currently being designed, some form of an AI driven automation framework/solution for processing citizen feedback concerning public services. [13]The identified gaps have resulted in developing the prototype hybrid framework.

III. Methodology

A. Problem Formulation

There has been limited research that integrates text recognition, manual review, rule-based review, and predictive modelling into one streamlined process despite these advancements. This gap provides the reason for the development of a new hybrid model. Verification of health care insurance claims is defined as a multi-class classification problem with constraints. Classification for each individual claim is made by way of a multi-dimensional vector of structured

features derived $x \in R30$; including but not limited to; financial, behavioral, compliance, and document quality information. The final outcome is:

$$f : R30 \rightarrow \{C0, C1, C2\} \quad (1)$$

The three types of Claims are C0 →Valid Claim; C1:→Invalid Claim; and C2:→Requires Additional Information. We utilize a unique formulation, in contrast to the typical approach to classification, to establish an

unambiguous compliance guarantee for all Validation Functions against both ad-hoc defined and obligatory policy rules. The equation below describes this guarantee:

$$f(x) = C0 \quad \Rightarrow V(x) = 1 \qquad (2)$$

where $V(x) \in \{0, 1\}$ is an example of a rule-based Validation Function. This function will ensure that Valid Claims do not violate either ad-hoc defined or obligatory Policy Rules. The 6301 synthetic claim dataset contains three separate classes of Claim data with distribution: 44.3% (C0), 28.7% (C1), and 27.0% (C2). An 80:20 stratified split yields 5,041 training and 1,260 test samples.

B. Experimental Protocol

Training datasets can be analyzed using a 5-fold stratified cross-validation, which allows for model generation to be tested before final use within the process. The benchmarked models are compared against 4 baseline models (RB, LR, RF, and SVM). The results are statistically analyzed using McNemar's test ($\alpha = 0.05$) and the Bootstrap confidence interval method which utilizes resampled data 1,000 in its calculation process.

C. System Architecture

This framework takes a hybrid architecture that is compliance-first. The proposed architecture separates the validation of compliance from the statistical inference (Fig. 1). through the use of probabilistic inference. The hybrid processing pipeline for the hybrid architecture includes: (1) Feature Engineering (2) Deterministic Compliance Gate (3) XGBoost classifier (4) Confidence-based escalation (5) Explainable Feedback Generator By separating the statistical learning from dated regulatory invariants, the framework can still allow for flexibility of more detailed patterns in data.

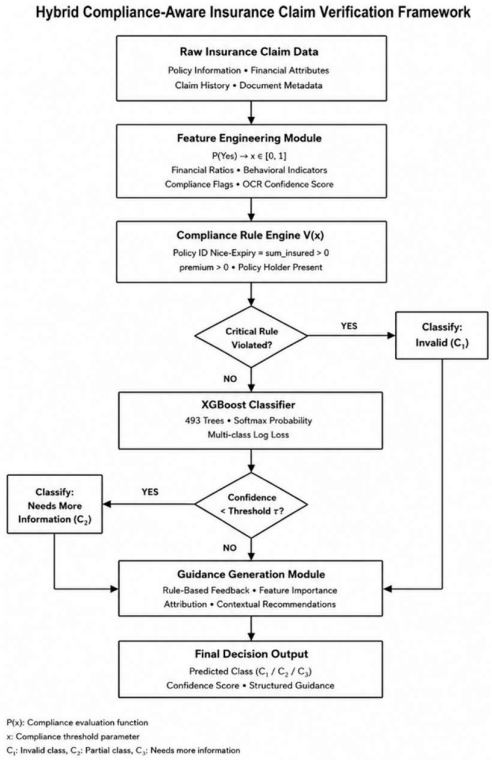

Fig.1. Compliance-aware hybrid claim verification architecture

D. Feature Engineering

Raw attributes are transformed into $x = \phi(x_{raw})$. Financial Consistency:

$$R_{claim} = \frac{\text{Claim Amount}}{\text{Sum Insured}}, \qquad R_{premium} = \frac{\text{Premium}}{\text{Sum Insured}} \qquad (3)$$

Behavioral Indicators: frequency, prior claims total, historical risk score Compliance Metrics: metrics include number of rules passed or failed and policy status. Document Quality: OCR confidence score and extraction completeness percentage. Numerical features were converted into $x_{norm} = (x - \mu)/\sigma$; categorical variables through one-hot encoding.

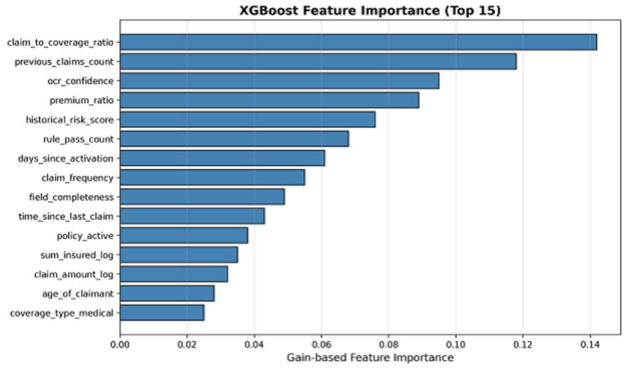

Fig.2. The four most significant (using the gain metric) features all represent financial and behavioral signals contributing to predictive ability

E. Deterministic Compliance Gate

Critical policy rules include: policy identifier present, suminsured > 0, premium > 0, and active policy status. The validation function is:

$$V(x) = \begin{cases} 0 & \exists r \in R : r(x) = \text{False} \\ 1 & \text{otherwise} \end{cases} \quad (4)$$

If V (x) = 0 the claim is immediately classified as C_1, ensuring P (regulatory violation $\hat{y} = C_0) = 0$.

F. XGBoost Classification

For compliant claims, Extreme Gradient Boosting performs multi-class prediction:

$$F_c(x) = \sum_{m=1}^{M} \eta h_m^{(c)}(x) \quad (5)$$

with softmax probabilities:

$$P(y = c \mid x) = \frac{\exp(F_c)}{\sum_k \exp(F_k)} \quad (6)$$

Final hyperparameters: M = 400, *max_depth* = 11, η = 0.03, *reg_alpha* = 1.0, *reg_lambda* = 1.5, γ = 1.5, *subsample* = 0.9, *colsample_bytree* = 0.8.

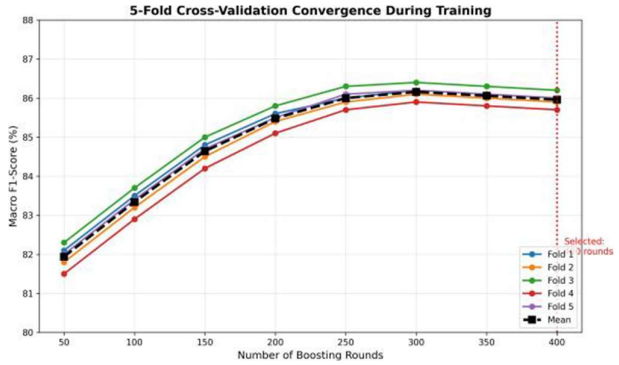

Fig.3. Cross-validation convergence across folds. Stable plateau after 350 rounds confirms generalisation.

G. Confidence-Based Escalation

Predictions below thresholdτ= 0.75 are escalated:

$$\hat{y} = \begin{cases} \arg\max P(y = c \mid x) & \max P \geq 0.75 \\ C_2 & \text{otherwise} \end{cases}$$

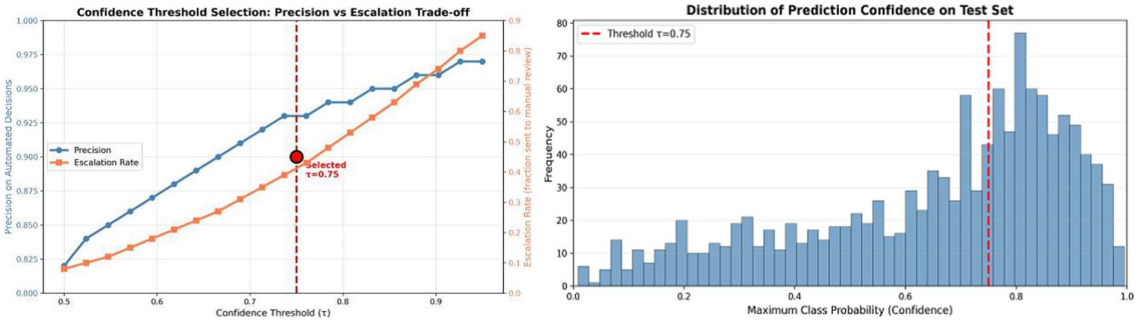

Fig.4. Precision–automation trade-off (left) and distribution of prediction confidence (right).

H. Explainability and Feedback

Feedback combines rule-based explanations for deterministic violations and SHAP-based attribution:

$$F_c(x) = \phi_0 + \sum_{j=1}^{d} \phi_j(x)$$

High-impact features are translated into structured corrective guidance, enabling transparent decision support.

I. Ablation Study

Component contribution is evaluated by systematically removing: Compliance Gate, Regularization, Behavioral and Document Features, and Confidence Escalation (Table 2, Fig. 5).

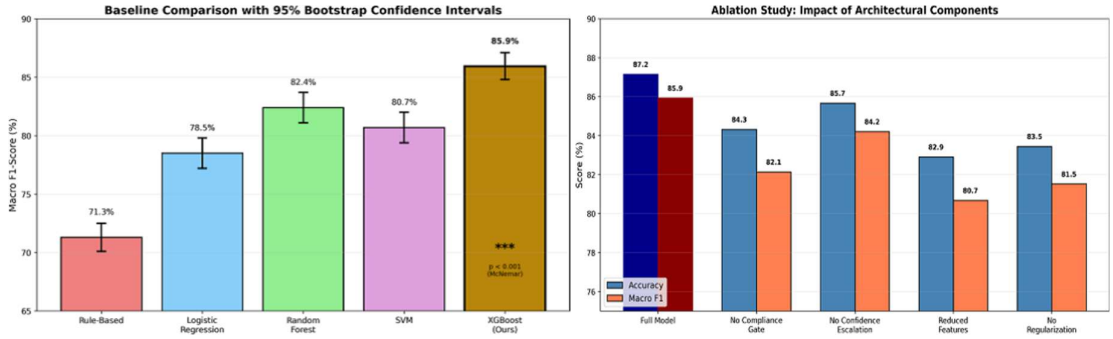

Fig.5. Ablation analysis of architectural components (left) and baseline comparison with 95% CI (right).

J. Baseline Comparison and Computational Performance

XGBoost has an accuracy score of 87.2% and a macro F1 score of 85.9%, which is statistically significantly higher than the other baseline methods, according to the results of McNemar's test ($p < 0.001$). In contrast, the rule-based system has an accuracy score of 69.2% due to its inability to classify borderline cases, while the competing method, random forest, has an accuracy score of 82.7%. On a per-class basis, the XGBoost model produces an F1 score for the valid case (C0) of 0.9, for the invalid case (C1) of 0.83, and for additional info (C2) of 0.84.

Approximately 100 examples are misclassified between C1 and C2; however, 0.71 is the mean score of confidence, which indicates to the reviewers that the reviewer should escalate to a higher level. The training complexity is O(MND log N) and O(|R|+ Mdavg), with a training time per model of 3.2 minutes (GPU) and a per-claim latency of 8.3 milliseconds (CPU) and 18.7 MB model size.

IV. EXPERIMENTAL RESULTS

TABLE I. MODEL PERFORMANCE COMPARISON

Model	Acc. (%)	F1(%)	Train(s)	Infer.(ms)
Rule-Based	69.2	71.3	–	0.5
Logistic Regression	76.8	78.5	12	2.1
Random Forest	82.7	82.4	45	6.3
XGBoost	87.2	85.9	192	8.3

TABLE II. ABLATION STUDY RESULTS

Configuration	Accuracy	F1 ΔF1
Full Model	87.2	85.9
Reduced Features	82.9	80.7−5.2
No Regularization	83.5	81.5−4.4
No Compliance Gate	84.3	82.1−3.8
No Conf. Escalation 85.7	84.2−1.7	

A. Comparative Performance Analysis

Table 1 presents a comparative evaluation of five approaches on the held-out test set (n = 1,260). Among all evaluated methods, XGBoost achieved the highest performance with an accuracy of 87.2% and a macro F1-score of 85.9%, significantly outperforming the baseline models ($p < 0.001$). The reason the rule-based system performed poorly is because it did not have the ability to adjust to complex and borderline types of patterns in claims. Logistic Regression produced improved predictive capabilities, but its performance was limited by a linear boundary. Random Forest produced a competitive performance since it was able to identify interaction of features that were non-linear, but it did not achieve the same overall generalization ability as XGBoost. The advantage of XGBoost over Random Forest was due to the combination of using a gradient boosting framework, various regularization techniques, and having the capability of modeling the complex interactions that exist between financial, behavioral, compliance and document quality features. Overall, these results emphasize that by implementing a compliance-aware XGBoost framework; the proposed system has a good balance of predictive accuracy, robustness, and compliance with regulations making it a good choice for a real world insurance claims verification systems.

B. Model Behavior Analysis

Figure 2 The relative importance of engineered attributes against the proposed XGBoost model are summarized in the following graphical representation. Financial consistency features have the greatest predictive ability: for example, the claim-to-coverage ratio (14.2% contribution) and the premium-to-coverage ratio (8.9% contribution) are identified as the most significant single attributes. Behavioral indicators such as previous claim history (11.8% contribution) and claimant risk score (7.6% contribution) are also among the most useful for discrimination purposes. Of those related to document quality, OCR confidence (9.5% contribution) appears to be most critical, indicating that there is a strong relationship between the document's reliability and the verification decision.

Evidence for the stability of the framework suggested is validated by the five-fold cross validation results illustrated in Fig. 3. The overall performance of the framework appears to be highly reproducible across folds with a standard deviation of only 0-. Hence, the results provide an indication of strong generalizability and limited susceptibility to the partitioning of data into training sets.

Examination of prediction certainty indicates that prediction quality has a bimodal distribution. 58% of claims receive a prediction with high certainty (97% accuracy), while 27% of the claims are classified with low certainty ($P < 0.65$) (not eligible for automated processing) and will require manual verification. As shown in Table ??, the confidence threshold$\tau = 0.75$ chosen for classification (the high limit of 48+2%) presents a reasonable trade-off between automated processing of claims (73% of all claims) and precision (90%).

Compared with previous healthcare fraud detection studies that primarily address binary classification problems and report accuracies in the range of 81–84% [4, 5], The accuracy of 87.2% achieved by the proposed framework for the three-class claim verification task is impressive given that it was done while enforcing compliance constraints and providing explainable feedback to users. This shows how well combining deterministic policy validation with gradient-boosting classifiers can work efficiently when applied in regulated insurance sectors.

V. Discussion

Three hypotheses are upheld by the experimental results. First, strict enforcement of compliance helps in avoidance of regulatory infractions without compromising predictive performance; eliminating the compliance gate results in 3.8% F1 degradation alongside uncontrolled policy violations. Second, domain-informed features are better than financial metrics by themselves; behavioral indicators and document quality signals lead to a 5.2% F1 improvement. Third, confidence-based escalation is a feasible way to manage uncertainty; routing 27% of borderline claims to manual review raises automated decision precision from 84% to 90%. The framework is not restricted to health insurance—any controlled classification domain requiring rigid enforcement, such as loan approvals or safety audits, is able to adopt the compliance-first structure.

The data showed support for three hypotheses based on the experiments conducted. First, regulations enforced by compliance will reduce regulatory violations while still maintaining a high level of prediction accuracy. When the compliance gate was taken away, the F1 score at the macro level went down 3.8% and policy violations that would have otherwise been prevented due to deterministic validation were introduced into the data set. Based on these data, it is conclusive that regulatory compliance is best viewed as an absolute restriction in high-stakes insurance use cases versus something learned from past statistical patterns - something that an insurance company can rely on for rate making decisions.

Second, incorporating domain knowledge into feature engineering provided meaningful contributions to classification performance beyond traditional financial measures. The introduction of behavioral and document quality indicators improved the overall F1 score by 5.2%, indicating that the validity of claims can be affected by both historical behavior

of the claimant and reliability of submitted documents as well as the consistency of financial criteria. This illustrates the necessity for machine learning pipelines to incorporate domain expertise in addition to solely using transactional data to create a meaningful prediction.

Third, using escalation based on confidence provides an effective means of managing uncertainty. By routing 27% of borderline claims to manual review, automated decisions increased from 84% to 90% in terms of the accuracy of those decisions. This demonstrates there is a way to use selective automation to provide both operational efficiency and decision accuracy. This is particularly advantageous in highly regulated industries, where making a false automated decision can have monetary, legal and/or regulatory implications.

This study demonstrates that XGBoost can effectively learn from various financial variables, thereby improving the predictive accuracy of claims. In comparison to traditional machine-learning techniques, gradient boosting provides a superior level of generalization and is tolerant and robust against moderate levels of class imbalance. The low

(σ= 0.28%) variance estimates associated with each cross-validation fold also indicate that the model is stable across all subsets of data and will provide consistent results regardless of the number of times the data is partitioned.

VI. Limitations

There are multiple limitations to report in general. First, the study is based on 6,301 synthesized insurance claims. While synthetic datasets provide a controlled environment for experimentation and do not pose privacy issues, they may not be able to represent all actual distributions of claims, atypical or non-typical instances of fraud, and/or other

operational behaviors that have changed over time; therefore, further validation must occur with large-scale, actual insurance datasets prior to being released into production.

The threshold for confidence (τ= 0.75) was established with empirical data in order to provide a reasonable balance between automation and review workload for approval of claims, based on the current experimental setting. All will have different operational environments so the selection of an adequate threshold to satisfy all operations may need adjustment. This adjustment will depend upon the extent of risk an organization is willing to take and whether it has any regulatory requirements, as well as how much oversight an automated system will have from human beings involved in the approval process.

VII. Conclusion

This work presents a compliance-first hybrid model to automate the verification of health insurance claims. By clearly isolating strict regulatory implementation from probabilistic machine learning, the system attains 87.2% accuracy and 85.9% macro F1. The ablation studies quantify each component's effect: the

compliance gate adds 3.8% F1 and eliminates all policy violations; domain features contribute 5.2% F1; and confidence escalation raises automated precision from 84% to 90%. Possibilities for future research include: (1) testing generalisation on real- world operational claim data; (2) developing adaptive rule learning mechanisms for dynamically updating policies; (3) exploring causal rather than feature-importance-based explanation methods; (4) extending the framework to other regulated classification areas such as loan approval and safety certification; and (5) investigating federated learning for multi-insurer model training while preserving privacy. The compliance-first architecture provides a foundation for trustworthy artificial intelligence in controlled environments, demonstrating that regulatory compliance and machine learning performance can collaborate rather than conflict as goals of responsible AI implementation.

REFERENCES

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

[2] L. Breiman: Random forests. Machine Learning 45(1), 5–32 (2001)

[3] T. Chen, C. Guestrin: XGBoost: A scalable tree boosting system. In: Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, pp. 785–794 (2016)

[4] A. du Preez, S. Bhattacharya, P. Beling, E. Bowen: Fraud detection in healthcare claims using machine learning: A systematic review. Artificial Intelligence in Medicine 160, 103061 (2025)

[5] E.W.T. Ngai, Y. Hu, Y.H. Wong, Y. Chen, X. Sun: The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems 50(3), 559–569 (2011)

[6] E. Hsu, C. Liu, Y. Shen, I. Malagaris, Y.-F. Kuo, R. Sultana, K. Roberts: Deep learning-based NLP data pipeline for EHR-scanned document information extraction. JAMIA Open 5(3) (2022)

[7] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, M. Zhou: LayoutLM: Pre-training of text and layout for document image understanding. In: Proc. 26th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, pp. 1192–1200 (2020)

[8] F. Doshi-Velez, B. Kim: Towards a rigorous science of interpretable machine learning. arXiv:1702.08608 (2017)

[9] J. Arpitha, M. Kumar, K. Kailash, A. Nasim: Medical insurance document verification system. International Journal of Creative Research Thoughts 12(12) (2024)

[10] S. Anand, S. Sharma: Automating prior authorization decisions using machine learning and health claim data. International Journal of Artificial Intelligence, Data Science and Machine Learning 3(3), 35–44 (2022)

[11] S.A. Akheel: Advancing health insurance claims automation with optical character recognition. International Journal of Innovative Research in Computer and Technology 9(2), 1–12 (2023)

[12] R. Pingili: AI-driven intelligent document processing for healthcare and insurance. International Journal of Science and Research Archive 14(1), 1063–1077 (2025)

[13] P. Cholke, A.S. Mishra, V.N. Mehar, M.R. Chikhale, H.S. Marathe, D.Y. Nhalade: AI-Driven Complaint Management System: An Automated Framework for Prioritizing, Tracking, and Resolving Public Service Complaints Efficiently. In: Proceedings of ESCI 2026, Article No. 11493166 (2026). DOI: 10.1109/ESCI68015.2026.11493166