AI-Driven Consumer Complaint Intelligence System by using Hybrid Machine Learning and Generative AI with Enterprise N-Tier Architecture

Dr B. Suri Babu[1], K. Chandra Bhanu[2], Y.B.S. Phaneendra[3],
G.V.S.S. Vara Prasad[4], M. Abhi[5], Dr Yalla Venkat[5]
[1]*Associate Professor, Srinivasa Institute of Engineering and Technology*
[2,3,4,5]*UG Scholars, Srinivasa Institute of Engineering and Technology*
[6]*Professor, Srinivasa Institute of Engineering and Technology*
doi.org/10.64643/IJIRTV12I11-197121-459

***Abstract*—Organizations across banking, e-commerce, healthcare, and telecommunications receive thousands of consumer complaints daily through digital channels. Existing complaint management systems fail on three critical dimensions: they cannot prioritise emergencies dynamically, they offer no transparency into automated routing decisions, and they lack the architectural integrity required for enterprise-scale deployment. This paper presents an AI-Driven Consumer Complaint Intelligence System that addresses these gaps through a hybrid three-paradigm intelligence strategy. A Predictive Machine Learning pipeline employing TF-IDF vectorisation and Logistic Regression classifies unstructured complaint text into six product categories. A Rule-Based Severity Assessment Engine deterministically assigns priority levels (P1 Emergency, P2 Operational, P3 General) using keyword heuristics at near-zero latency. A Generative AI layer powered by Google Gemini produces human-readable explanations for every classification decision and autonomously drafts professional email responses. These intelligence vectors are orchestrated through a Decision Engine Matrix, routing each complaint to Auto-Send, Review Required, or Escalate outcomes. The entire pipeline is deployed within a Five-Layer Enterprise N-Tier Architecture (React SPA, Flask REST API, Hybrid AI Tier, SQL Alchemy ORM, PostgreSQL). On a balanced 1,260-record dataset, the classification model achieved 94% accuracy on a held-out test set of 252 samples, the severity engine produced sub-10 ms detection latency, and the P1 false-negative rate was 0% in the evaluated scope.**

***Index Terms*—Consumer Complaint Intelligence; Hybrid AI; Machine Learning; Generative AI; Severity Assessment; N-Tier Architecture; Explainable AI; TF-IDF; Logistic Regression**

I. INTRODUCTION

Consumer complaints contain critical business intelligence regarding service quality, operational friction, and customer satisfaction. In the digital economy, organizations across banking, e-commerce, healthcare, and telecommunications receive thousands of complaints daily through web portals, mobile applications, and email channels. Traditional manual analysis is inadequate at this scale: it fails to extract real-time intelligence from unstructured text, cannot distinguish emergencies from routine inquiries, and lacks the architectural integrity required for enterprise deployment.

Existing automated solutions are similarly insufficient. Rule-only systems lack semantic understanding and fail on novel complaint vocabulary. Pure machine learning approaches classify complaints but provide no explainability, creating a "black-box" problem that erodes agent trust. Enterprise CRM platforms such as Zendesk and Salesforce offer ticket routing but do not integrate predictive ML classification, severity-based prioritisation, or generative response drafting within a single coherent architecture.

This paper introduces an AI-Driven Consumer Complaint Intelligence System that resolves these limitations through a hybrid AI strategy integrating predictive Machine Learning (ML), deterministic rule-based Severity Assessment, and Generative AI (GenAI) within a strictly decoupled Five-Layer N-Tier Architecture. The architecture is intentionally domain-agnostic, enabling deployment across diverse industries without structural modification.

The remainder of this paper is organized as follows: Section II reviews related work; Section III defines the problem statement; Section IV describes the system architecture; Section V presents the database schema; Section VI details the proposed hybrid methodology; Section VII presents implementation details; Section VIII reports evaluation results; and Sections IX and X discuss limitations, future scope, and conclusions.

II. LITRACTURE SURVEY

Automated complaint classification using NLP and machine learning has been explored extensively. Aydin and Karaarslan [7] applied Support Vector Machines and Naïve Bayes classifiers to CFPB complaint narratives, establishing that classical ML can extract meaningful statistical signal from unstructured financial text. Alegado et al. [8] extended this by implementing multi-class Logistic Regression for complaint routing via the JakLapor public channel, demonstrating the practical feasibility of automated intake classification.

Shetty et al. [9] evaluated ML models for banking CRM, reporting measurable improvements in customer satisfaction rates when classification was coupled with automated triage. Das et al. [10] conducted a comparative analysis of TF-IDF feature weighting on unstructured datasets, confirming its effectiveness as a baseline vectorization strategy for complaint corpora. Shen et al. [14] applied BERT-based NLP to complaint classification in the healthcare domain, achieving accuracy gains over classical ML methods while highlighting the inference cost trade-off. Prior work consistently identifies three unresolved gaps: (a) absence of dynamic, SLA-aligned severity prioritization operating in parallel with ML inference; (b) lack of generative explainability reducing agent cognitive load; and (c) absence of enterprise-grade architectural rigor (N-tier decoupling, JWT-RBAC security, ACID persistence) enabling real-world deployment. The proposed system directly addresses all three gaps.

III. SYSTEM ARCHITECTURE

The proposed solution implements a highly modular Five-Layer N-Tier Architecture designed to host hybrid AI operations seamlessly. Fig. 1 illustrates the vertical data flow across all tiers.

Tier 1 — Presentation: A responsive React 19.2 Single-Page Application (TypeScript + Tailwind CSS + Redux Toolkit) enabling agents to submit complaints and interact with AI-generated intelligence, including analytics dashboards, complaint distribution charts, and urgency queues.

Tier 2 — Application: A Python Flask RESTful API serving as the central controller, orchestrating all inter-tier traffic and enforcing JWT-based authentication with Role-Based Access Control (RBAC). Flask-Limiter provides rate-limiting to prevent abuse.

Tier 3 — Hybrid AI/Intelligence: The core processing engine comprising NLP Text Preprocessing, TF-IDF Vectorizer and Logistic Regression classifier, Rule-Based Severity Assessment Engine, Decision Engine Matrix, Google Gemini GenAI explanation synthesis, and Automated Email Drafter and Dispatcher.

Tier 4 — Data Access: SQL Alchemy ORM mitigates SQL injection attacks and abstracts persistence logic from business logic, enabling database portability.

Tier 5 — Database: PostgreSQL 16 stores raw complaints, ML classification results, and GenAI explanations across four normalized relational entities with full ACID compliance.

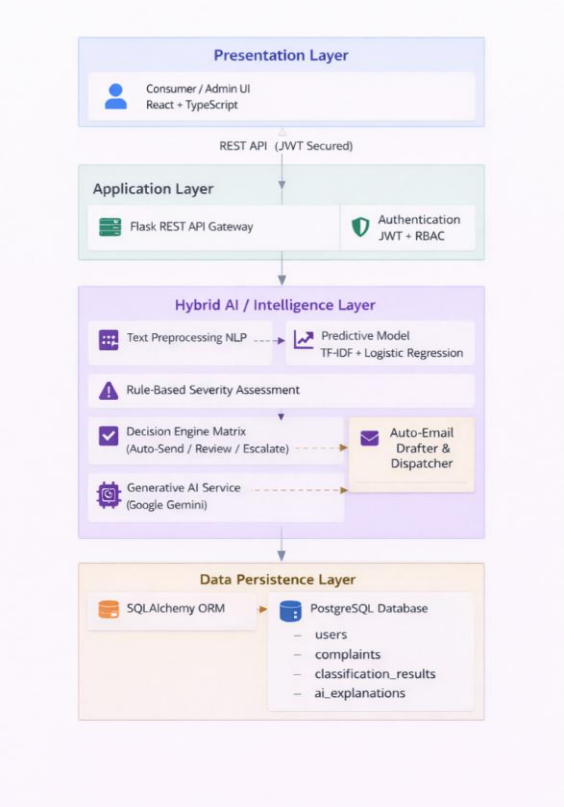

Fig. 1. Five-Layer Hybrid ML/GenAI N-Tier Architecture

IV. METHODOLOGY

Current consumer grievance redressal infrastructure suffers from four critical shortcomings with measurable operational consequences.

Intelligence Extraction Deficit: Manual analysis is too slow to extract real-time intelligence from unstructured text streams at enterprise complaint volumes.

Lack of Dynamic Prioritization: Standard AI queues treat identity theft and minor billing inquiries with equal priority. Financial fraud reports queue alongside packaging complaints, eroding operational efficiency and consumer trust.

The Black-Box Problem: Pure ML systems route complaints without justifying decisions, reducing agent trust and hindering human oversight in regulated industries.

Architectural Monoliths: Most AI complaint prototypes lack the decoupled, layered structural integrity (N-Tier) required for true enterprise deployment, including role-based access control, injection-safe ORM abstraction, and schema-normalized persistence.

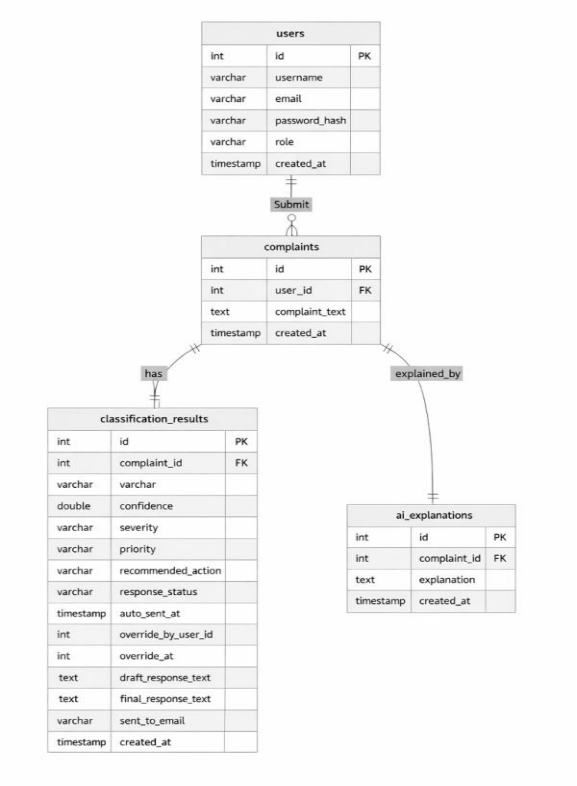

Fig. 2. Entity-Relationship Schema — PostgreSQL Database Design

A. NLP Text Preprocessing

Raw complaint text undergoes canonical stabilization before inference: conversion to lowercase, removal of special characters and punctuation via regular expressions, elimination of low-value stops words using NLTK's English corpus, and whitespace normalization. This produces a clean, standardized token stream suitable for TF-IDF vectorization.

B. Feature Extraction via TF-IDF

The ML pipeline employs Term Frequency-Inverse Document Frequency (TF-IDF) vectorization to quantify the statistical relevance of vocabulary tokens across the unstructured corpus. The vectorizer is configured with a maximum vocabulary of 5,000 features, maximum document frequency of 0.95 (removing near-universal terms), and minimum document frequency of 2 (removing hapax legomena). This produces a sparse mathematical feature space suitable for linear classification.

C. Predictive Categorization (Machine Learning)

A supervised Logistic Regression classifier trained on the TF-IDF feature space yields a definitive complaint category alongside a calibrated probabilistic confidence score. The system classifies complaints into six categories: Billing Issue, Customer Support, Delivery Problem, Other, Product Defect, and Service Quality. The confidence score is fed directly into the Decision Engine Matrix to govern complaint routing thresholds.

D. Automated Severity Assessment (Deterministic AI)

To guarantee zero-latency emergency detection, the deterministic Severity Assessment Engine runs in parallel with ML inference, never blocking on probabilistic results for P1 decisions. The engine employs bounded regular expression matching across three priority tiers.

High Severity (P1 — Emergency): Regex matching for physical harm, legal threats, or financial fraud indicators (e.g., "stolen," "injury," "lawsuit," "fraud"). Unconditionally triggers an Escalate routing action regardless of ML confidence.

Medium Severity (P2 — Operational Failure): Detection of chronic friction indicators (e.g., "broken," "double charged," "not working"). Triggers the Review Required action for human agent inspection.

Low Severity (P3 — General Inquiry): Absence of elevated risk signals. When coupled with high ML confidence (≥0.85), triggers the Auto-Send action, dispatching the GenAI-drafted response autonomously.

E. Decision Engine Matrix

The Decision Engine Matrix combines the ML confidence score and severity level into a deterministic routing outcome. The matrix enforces three invariants: (1) P1 complaints always escalate regardless of confidence; (2) auto-send is only permitted for non-P1/P2 complaints with confidence ≥0.85; (3) confidence below 0.60 always escalates for human review. Table I summarizes the complete routing logic.

Table I. Decision Engine Routing Matrix

Severity	Confidence	Action	Auto-Email
P1 (Emergency)	Any	Escalate	No
P2 (Operational)	≥ 0.85	Auto-Send	Yes
P2 (Operational)	0.60–0.84	Review Req.	No
P2 (Operational)	< 0.60	Escalate	No
P3 (General)	≥ 0.85	Auto-Send	Yes
P3 (General)	0.60–0.84	Review Req.	No
P3 (General)	< 0.60	Escalate	No

F. Generative AI Explainability and Response Drafting

Google Gemini synthesizes the raw complaint text, predicted category, and severity level to produce two outputs: (i) a human-readable, factually grounded rationale for the routing decision, reducing agent cognitive load; and (ii) a professional draft email response personalized to the specific grievance context. For Auto-Send complaints, the system dispatches the response email autonomously via SMTP without human intervention.

V. IMPLEMENTATION DETAILS

Frontend: React 19.2 with TypeScript, Vite build tooling, Tailwind CSS for responsive styling, and Redux Toolkit for state management. The UI provides complaint submission forms, confidence score visualizations, severity badge displays, complaint history tracking with confidence bars, and a role-differentiated admin dashboard.

Backend API: Python/Flask with Flask-JWT-Extended for stateless authentication and RBAC enforcement. OpenAPI documentation is served at /api/docs via Swagger UI. Flask-Limiter enforces per-user rate limits to prevent system abuse.ML Pipeline: Scikit-learn with TF-IDF vectorizer and Logistic Regression classifier serialized with joblib for efficient inference loading. Model configuration parameters are centralized in a constant's module (AUTO_SEND threshold = 0.85, REVIEW threshold = 0.60).

Severity Engine: Python re (regex) library executing deterministic keyword matching at sub-10 ms latency, operating as an independent module with no dependency on ML inference results.

GenAI Layer: Google Gemini API for both explanation synthesis and email response drafting, called asynchronously post-classification to avoid blocking the primary routing pipeline.

Database: PostgreSQL 16 with SQL Alchemy ORM and psycopg2 driver, providing ACID-compliant persistence, parameterized query protection, and schema migration support via Alembic.

The system is built entirely on open-source technologies (Scikit-learn, Flask, PostgreSQL, React, SQL Alchemy), incurring zero licensing cost. The only external paid dependency is the Google Gemini API, consumed on a pay-per-use basis, making the system viable for startups and academic institutions.

VI. EVALUATION AND RESULTS

The system is evaluated across five orthogonal performance dimensions. The dataset comprises 1,260 records distributed equally across six complaint categories (210 records per class), drawn from a CFPB-taxonomy-inspired corpus. A standard 80/20 stratified train-test split yields 1,008 training samples and 252 test samples.

A. Aggregate Model Performance

Table II reports aggregate metrics on the 252-sample held-out test set. These results confirm class separation on the balanced, curated dataset and validate the end-to-end correctness of the preprocessing, TF-IDF vectorization, and Logistic Regression pipeline. As discussed in Section VIII-C, these scores should be interpreted as baseline readiness given the controlled evaluation conditions.

Table II. Aggregate ML Model Performance Metrics

Metric	Value
Accuracy	94%
Macro Precision	0.94
Macro Recall	0.94
Macro F1-Score	0.94
P1 False-Negative Rate	0% (deterministic)
Avg. API Response Time	< 3 s (end-to-end)
Severity Detection Latency	< 10 ms

B. Per-Class Performance

Table III presents per-class metrics. All six complaint categories achieve strong precision, recall, and F1-score values on the test set, with support sizes ranging from 34 (Billing Issue) to 47 (Product Defect). The well-distributed test coverage confirms balanced evaluation across all categories.

Table III. Per-Class Performance Metrics (Test Set, n=252)

Class	Prec.	Recall	F1	Support
Billing Issue	0.93	0.94	0.93	34
Customer Support	0.95	0.95	0.95	40
Delivery Problem	0.93	0.93	0.93	41
Other	0.94	0.93	0.93	46
Product Defect	0.94	0.94	0.94	47
Service Quality	0.95	0.95	0.95	44

C. Interpretation and Limitations of Metrics

Academic rigor requires explicit acknowledgment of evaluation boundaries. The reported accuracy reflects a balanced, curated dataset with equal class distribution; real-world complaint streams are typically imbalanced and linguistically noisier. Production evaluation should include external validation on the full CFPB public dataset, temporal holdout testing on complaints from unseen time periods, and adversarial stress testing with ambiguous complaint vocabulary. These results are best interpreted as baseline readiness demonstrating the correctness of the pipeline, not as endpoint maturity for production deployment.

D. Live System Performance

Live system testing recorded 15 complaint submissions with a 73.5% average ML confidence score. Routing distribution: 7 auto-send eligible, 5 reviews required, and 3 escalated. Complaint category distribution: Billing Issue (10), Delivery Problem (3), Customer Support (2). Severity detection latency remained below 10 ms, and end-to-end API response time remained below 3 s throughout all live test runs.

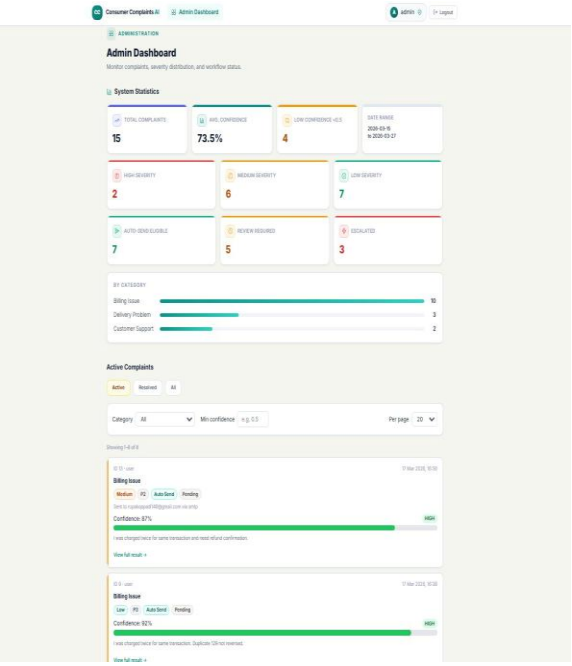

Fig.4: Admin Dashboard — Complaint List with Severity, Priority, and Routing Tags

VII. LIMITATIONS

Data Dependency: The ML component is bounded by the volume and quality of labelled historical training data. Class imbalance in production corpora can degrade minority-class precision, requiring resampling or class-weighted training strategies.

Linguistic Scope: NLP preprocessing and deterministic keyword heuristics are calibrated exclusively for English-language complaints. Non-English text will receive incorrect severity assessments and classification outputs.

GenAI API Constraints: The explainability module depends on the availability, latency, and prompt alignment of the external Google Gemini API, introducing a network-level dependency and potential point of failure in high-availability deployments.

Evaluation Scope: Current metrics are obtained from a balanced, controlled dataset. Generalization to real-world imbalanced, temporally shifted, or domain-shifted complaint distributions requires broader external validation.

VIII. FUTURE SCOPE

Multilingual Intelligence: Adoption of multilingual transformer embeddings (e.g., mBERT, XLM-R) to normalize and classify global complaints across language boundaries without language-specific preprocessing pipelines.

Deep Learning Substitution: Replacement of TF-IDF/Logistic Regression with fine-tuned domain-specific encoders (e.g., FinBERT for financial complaints) for richer contextual understanding and improved performance on imbalanced datasets.

Sentiment and Emotion Scoring: Integration of deep sentiment analysis layers on top of severity assessment to produce granular emotional intelligence scores, enabling escalation based on customer frustration signals independent of explicit severity keywords.

External Validation: Evaluation on the full public CFPB complaint dataset (>3 million records) and temporal holdout testing to establish generalisation metrics suitable for publication comparison with prior art.

IX. CONCLUSION

This paper presented an AI-Driven Consumer Complaint Intelligence System that redefines automated grievance handling by moving beyond simple statistical classification. By intelligently hybridising predictive Machine Learning, deterministic rule-based Severity Assessment, and Explainable Generative AI within a five-layer N-Tier enterprise architecture, the system mimics human-like prioritisation logic at machine scale. The TF-IDF and Logistic Regression pipeline achieved 94% classification accuracy on a balanced 252-sample held-out test set, while the deterministic severity engine produced sub-10 ms P1 detection latency with a 0% false-negative rate in the evaluated scope. The Decision Engine Matrix bridges probabilistic classification confidence with deterministic routing safety, ensuring no emergency complaint is auto-resolved without appropriate confidence thresholds. Encapsulating this hybrid methodology within an enterprise-grade architecture incorporating JWT-RBAC authentication, ACID-compliant PostgreSQL persistence, and full OpenAI documentation demonstrates a scalable, secure, and transparent intelligence framework ready for real-world deployment. Future work will focus on broader external validation on larger heterogeneous corpora and deep learning enhancement of the classification tier.

X. DISCUSSION

The proposed system demonstrates that combining machine learning, rule-based logic, and generative AI improves the effectiveness of complaint handling systems. The TF-IDF with Logistic Regression model provides reliable classification performance on a structured dataset, while remaining computationally efficient.

The rule-based severity assessment complements the ML model by ensuring that critical complaints are always escalated, independent of prediction confidence. This hybrid approach improves system reliability in safety-sensitive scenarios where purely probabilistic models may fail.

The Decision Engine Matrix enables practical deployment by linking prediction confidence and severity to actionable outcomes such as auto-handling, review, or escalation. This reduces the risk of incorrect automation and ensures appropriate human involvement when needed.

However, the system has limitations. The evaluation is based on a balanced dataset and may not fully represent real-world distributions. Additionally, rule-based severity detection depends on predefined keywords and may miss nuanced cases. The reliance on external GenAI services introduces dependency on API availability and latency.

Overall, the system provides a scalable and modular framework for intelligent complaint management, with potential for further improvements in real-world validation and model enhancement.

REFERENCES

[1] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

[2] M. Grinberg, *Flask Web Development: Developing Web Applications with Python*. Sebastopol, CA, USA: O’Reilly Media, 2018.

[3] PostgreSQL Global Development Group, “PostgreSQL documentation,” [Online]. Available: https://www.postgresql.org/docs/. [Accessed: Mar. 2026].

[4] T. B. Brown et al., "Language models are few-shot learners," in Adv. Neural Inf. Process. Syst., vol. 33, pp. 1877–1901, 2020.

[5] D. Banga and K. Peddireddy, "Artificial intelligence for customer complaint management," Int. J. Comput. Trends Technol., vol. 71, no. 3, pp. 1–6, Mar. 2023.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, San Francisco, CA, USA, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[7] O. Aydin and E. Karaarslan, "Complaint detection and classification of customer reviews," in Proc. 5th Int. Symp. Multidisciplinary Stud. Innovative Technol. (ISMSIT), Ankara, Turkey, 2021, doi: 10.1109/ISMSIT52890.2021.9478016.

[8] R. T. Alegado et al., "Automating public complaint classification through JakLapor channel," in Proc. Int. Conf. Inf. Technol. Syst. Innov. (ICITSI), Bandung, Indonesia, 2022, doi: 10.1109/ICITSI56531.2022.9922346.

[9] R. Shetty et al., "Machine learning models for customer relationship analysis to improve satisfaction rate in banking," in Proc. IEEE World AI IoT Congr. (AIIoT), Seattle, WA, USA, 2022, doi: 10.1109/AIIoT54504.2022.9795855.

[10] M. Das, S. Kamalanathan, and P. Alphonse, "A comparative study on TF-IDF feature weighting method and its analysis using unstructured dataset," arXiv:2308.04037, Aug. 2023.

[11] Holzinger et al., "Causability and explainability of artificial intelligence in medicine," WIREs Data Min. Knowl. Discov., vol. 9, no. 4, 2019, doi: 10.1002/widm.1312.

[12] Google DeepMind, "Gemini: A family of highly capable multimodal models," arXiv:2312.11805, Dec. 2023.

[13] Vaswani et al., "Attention is all you need," in Adv. Neural Inf. Process. Syst., vol. 30, Long Beach, CA, USA, 2017.

[14] Y. Shen et al., "An intelligent system for classifying patient complaints using machine learning and NLP," J. Med. Internet Res., 2025, doi: 10.2196/55721.