OPEN ACCESS

Haya: The Saudi Journal of Life Sciences
Abbreviated Key Title: Haya Saudi J Life Sci
ISSN 2415-623X (Print) | ISSN 2415-6221 (Online)
Scholars Middle East Publishers, Dubai, United Arab Emirates
Journal homepage: https://saudijournals.com

Original Research Article

A Multidimensional, Transformer-Based Framework for Predicting Physician Popularity on Online Health Platforms

Muhammad Umer Imran[1,2]*, Syed Jaffar Raza[2], Song Yiying[3,4], Syed Nouman Ali Shah[5], Syed Danyal Ali Naqvi[6], Asad Rehman[7]

[1]Clinical Medicine, Nanchang University, Nanchang, China
[2]Machine Learning Research Associate, 3Dex Inc., Irvine, CA 92617, USA
[3]Data Science and Big Data Technology, Shenyang University, Shenyang, China
[4]Data Science Associate, Shanghai Jingci Technology Co., Ltd., Shanghai, China
[5]School of Computer and Information Technology, Beaconhouse National University, Lahore, Punjab, Pakistan
[6]Department of Computer Science, COMSATS University, Islamabad, Pakistan
[7]Department of Computer Science, Virtual University, Islamabad, Pakistan

DOI: https://doi.org/10.36348/sjls.2025.v10i11.009 | **Received:** 04.11.2025 | **Accepted:** 27.12.2025 | **Published:** 31.12.2025

***Corresponding author:** Muhammad Umer Imran
Clinical Medicine, Nanchang University, Nanchang, China

Abstract

Digital health portals increasingly depend on highly "popular" physicians to anchor user traffic and drive revenue. Existing work, however, (i) conflates popularity with a single behavioural cue (consultation count) and (ii) relies on linear or shallow machine-learning models. We introduce PopNet, a hybrid TabTransformer + GRU that fuses demographic, behavioural, visual-cue and temporal-momentum signals to predict a composite Popularity Index (PopIdx) built from four pillars: demand, monetary appreciation, social proof and visibility. Across a five-fold group-wise cross-validation on 19 200 physician-quarter snapshots, PopNet attains MAE ≈ 0.091, beating ElasticNet by >40 %. Nevertheless, modern tree ensembles still edge it out (LightGBM MAE ≈ 0.046). Integrated-Gradient explanations and a feature-family ablation reveal platform visibility (*inv_rank*) as the single most important driver of popularity, followed by raw patient demand and monetary gifts. Fairness audits show a modest 0.006 PopIdx MAE gap between genders; a simple inverse-propensity re-weighting halves this gap with <0.002 performance loss. The study provides actionable levers for platform managers and a reusable, bias-audited modelling pipeline for future research.

Keywords: PopNet, Hybrid Deep Learning, TabTransformer, Gated Recurrent Unit (GRU), Integrated Gradients.

1 INTRODUCTION

1.1 Background

The proliferation of online health communities (OHCs) has redefined how patients access healthcare, enabling scalable and geographically agnostic interactions between patients and licensed physicians. These platforms serve as intermediaries, offering a hybrid of asynchronous (e.g., message-based) and synchronous (e.g., video or live chat) medical consultations (Peng *et al*., 2021). As the digital health landscape matures, platforms such as Haodf, Ping an Good Doctor, and WebMD have accumulated large user bases and vast repositories of physician–patient interactions. However, the long-term sustainability and profitability of such platforms do not solely depend on user engagement or satisfaction metrics but increasingly on the popularity of individual physicians.

Physician popularity, in this context, extends beyond clinical competence and encompasses a blend of platform visibility, social proof, patient loyalty, and perceived credibility. Popular doctors play a pivotal role in shaping patient traffic patterns. Their profiles are more frequently visited, their content is more widely shared, and they often command higher consultation fees and receive more patient gifts an online analogue of appreciation and loyalty (Wei & Hsu, 2022). These doctors also serve as informal influencers whose behavior and digital personas shape trust dynamics, satisfaction ratings, and even treatment choices within the community (Hsu *et al*., 2022).

From a business standpoint, popular physicians represent a form of "demand anchor" they stabilize traffic flows, enhance platform stickiness, and reduce

churn by ensuring that a core group of trusted experts remains constantly visible and accessible. In markets such as China, where competition among OHCs is intense, ensuring optimal physician visibility and forecasting potential "rising stars" are critical to platform orchestration and monetization strategies.

1.2 Research Gap

Despite the commercial and strategic significance of physician popularity, existing academic treatments often reduce this multi-dimensional construct to a single behavioral variable, such as consultation volume (Qin *et al*., 2022), rating scores, or follow counts (Gong *et al*., 2021). This oversimplification neglects the complex interactions among various signals that shape a physician's online presence and appeal. For instance, a doctor may achieve high visibility due to algorithmic ranking, not necessarily due to clinical excellence or patient satisfaction. Likewise, doctors who contribute high-quality educational content may enjoy high engagement but limited conversions to paid consultations.

From a methodological perspective, current literature is dominated by linear models (e.g., multiple regression, logistic regression) and classic machine learning ensembles such as random forests and gradient-boosted trees (Shah *et al*., 2022; Kaul *et al*., 2022). While these methods are robust and interpretable, they are inherently limited in capturing non-linear feature interactions, temporal dependencies, and visual-affective cues. Moreover, these models treat physician observations as independent and identically distributed (i.i.d.) samples, ignoring the temporal momentum of popularity how a doctor's performance or visibility in previous quarters influences future outcomes.

Another critical omission in prior work is the underutilisation of modern interface signals, such as facial expressions, profile credibility scores, and visual aesthetics. Research in social computing and trust modelling (Lyutkin *et al*., 2024) has shown that micro-expressions and profile cues (like smiling, perceived age, or facial symmetry) significantly impact how users perceive credibility and trustworthiness in digital environments. However, these cues are rarely integrated into predictive models of physician performance or popularity.

Given the above, there is a clear empirical and methodological gap: existing models fail to account for (a) the multi-pillar nature of popularity, (b) the temporal evolution of popularity trajectories, and (c) the visual-affective elements of digital health platforms. Furthermore, although transformer architectures have emerged as the state-of-the-art in vision, NLP, and tabular data modelling (Nerella *et al*., 2023; Vyas, 2024), their adoption in marketplace analytics especially in healthcare remains limited. This study aims to address these gaps.

1.3 Purpose and Significance

In response to these limitations, this paper introduces PopNet, a multidimensional, transformer-based framework designed to predict physician popularity within online health platforms. PopNet combines the structural strengths of the TabTransformer (which models high-order interactions across categorical codes) with a Gated Recurrent Unit (GRU) encoder that captures temporal sequences at the physician level. This hybrid architecture allows for simultaneous modelling of static demographics, behavioral indicators, time-series trends, and visual cues, offering a holistic view of physician popularity.

The study makes four key contributions:

1. **Methodological Advancement:** We develop a novel architecture that fuses transformer-based embeddings with GRU-encoded temporal trends. The architecture employs a multi-task learning setup, enabling the model to predict a composite Popularity Index (PopIdx) along with its four sub-pillars: demand, monetary appreciation, visibility, and social proof. The approach outperforms linear baselines and matches or exceeds state-of-the-art tree models in accuracy and generalizability.
2. **Empirical Benchmarking:** The model is trained and evaluated on a Chinese hospital-portal dataset, covering over 19,000 physician-quarter snapshots and 3,800 unique physicians. We benchmark PopNet against LightGBM, CatBoost, and ElasticNet, providing the most comprehensive evaluation to date of physician popularity forecasting.
3. **Interpretability and Transparency:** To avoid the "black-box" critique of deep learning, we incorporate Integrated Gradients (IG) for feature attribution and conduct a feature-family ablation analysis to assess the contribution of each modality (e.g., visibility, monetary signals, facial cues). This provides platform managers with actionable insights on which levers drive popularity and how to optimise platform design.
4. **Fairness and Bias Mitigation:** Given concerns around algorithmic fairness in healthcare, we implement a gender audit by comparing MAE (Mean Absolute Error) between male and female physicians. The model's fairness is evaluated under both standard training and inverse-propensity re-weighted training. We further test Group Distributionally Robust Optimization (GroupDRO) to assess trade-offs between fairness and predictive accuracy.

PopNet represents a methodologically novel and practically relevant solution to the challenge of popularity prediction in digital health ecosystems. By leveraging advanced representation learning and multi-task optimisation, it addresses limitations in previous studies while ensuring interpretability and equity critical

requirements for real-world adoption in healthcare platforms.

1.4 Research Questions

1. **RQ1:** How effectively can a transformer-augmented, multi-task neural architecture predict a composite Popularity Index relative to state-of-the-art tabular learners (CatBoost, LightGBM)?
2. **RQ2:** Which feature families demand, monetary appreciation, social proof, visibility, facial cues, or temporal momentum contribute most to model performance?
3. **RQ3:** Do popularity drivers differ across disease specialties and physician ranks, and how can the platform exploit these differences to balance patient load?
4. **RQ4 (Fairness):** Does the model exhibit disparate predictive error with respect to physician gender, and if so, which pipeline interventions reduce bias?

2 Literature Review and Conceptual Framework

2.1 Popularity in Online Health Communities (OHCs)

The concept of physician popularity in online health communities has garnered growing academic interest, especially as digital platforms become a primary channel for medical consultations and advice. Popularity in this context is not merely a function of medical expertise but emerges from a complex interplay of visibility, engagement, perceived trust, and interactive behavior. Several empirical studies underline the non-clinical determinants of physician popularity. For example, Ouyang *et al*., (2023) found that physicians who engage in free-knowledge sharing such as publishing educational articles or answering patient queries in public forums tend to attract significantly more visits and followers. This aligns with the broader theory of reciprocity in social exchanges, where informational generosity breeds trust and visibility.

Wei and Hsu (2022) further demonstrate that the thematic content of a physician's profile (e.g., focus on chronic diseases versus cosmetic procedures) influences the number and quality of patient ratings. Physicians discussing empathetic or family-oriented topics tend to score higher in ratings, independent of clinical effectiveness. The role of offline promotions also cannot be discounted. Hsu *et al*., (2022) show that participation in hospital-sponsored webinars or conferences often leads to short-term spikes in online consultations, suggesting that cross-channel visibility reinforces popularity. These findings collectively point to a multi-dimensional structure of popularity encompassing not just demand, but also social proof, monetary appreciation, and algorithmic ranking. Yet, most existing research tends to examine popularity using single outcome variables (e.g., consultation count or ratings), and often in isolation from modern machine learning approaches that can model high-order interactions across diverse feature sets.

2.2 Machine Learning in Health Portals

Predictive modelling in health portals has largely been dominated by tree-based ensemble models, especially gradient-boosted trees such as LightGBM, XGBoost, and CatBoost. These models are well-suited for tabular data and have demonstrated strong performance in diverse healthcare tasks, including patient triage, appointment no-show prediction, and chronic disease monitoring (Yang *et al*., 2024; Badawy *et al*., 2023). Their appeal lies in their ability to handle missing data, encode categorical variables, and offer post-hoc interpretability through feature importance scores. However, the limitations of such models are increasingly evident in tasks that require sequential understanding, multi-modal data fusion, or long-term temporal reasoning. This has led to a growing interest in deep tabular models that can embed high-dimensional categorical variables and learn complex interactions. Though still under-represented in healthcare applications, studies like Sumon *et al*., (2025) and Vyas (2024) report promising results using hybrid neural architectures for medical insurance fraud detection and patient churn forecasting, respectively.

Transformer architectures, originally designed for NLP, have now permeated nearly every domain of machine learning. Their ability to attend across token positions makes them ideal for tasks involving text, vision, and structured data. For instance, Acheampong *et al*., (2021) and Selvam *et al*., (2022) showcase the use of Vision Transformers (ViT) for interpreting radiographic images, while Raisi *et al*., (2021) applies BERT variants to electronic medical record (EMR) summarization. Recently, L'Heureux *et al*., (2022) and Zhang *et al*., (2022) have successfully applied transformers to time-series forecasting, including in finance and epidemiology. Despite these advances, transformers remain underutilized in marketplace analytics, especially for tasks involving the popularity prediction of agents (such as doctors) on platforms. This study bridges that gap by proposing a transformer-based architecture tailored to the heterogeneous and temporal nature of health portal data.

2.3 Explainable AI (XAI) in Healthcare

With the increasing complexity of predictive models in healthcare, explainability has become a crucial requirement. Stakeholders including patients, physicians, and platform regulators require transparency to ensure that algorithmic decisions are fair, understandable, and justifiable. Gradient-based attribution methods, such as Integrated Gradients (IG), have emerged as popular tools to trace the contribution of each input feature to a model's output (Wang *et al*., 2024). These methods are particularly well-suited for neural networks, providing fine-grained attributions that help identify dominant predictive signals.

Other approaches, such as self-explaining networks or attention-weight visualisations, also provide insight into what the model attends to during decision-making. Saraswat *et al.,* (2022) demonstrate the use of self-explaining networks in tabular datasets where interpretability is critical for auditability and regulatory compliance. PopNet incorporates Integrated Gradients for per-feature attribution and complements this with feature-family ablation, offering a dual-lens interpretability protocol that satisfies both research rigour and practitioner usability.

2.4 Fairness and Bias in Algorithmic Predictions

Fairness in AI systems especially those deployed in sensitive domains such as healthcare has emerged as a foundational pillar of ethical AI. Disparities in participation, exposure, and trust across demographic groups often result in predictive biases that can reinforce social inequities. Yin *et al*., (2022) argue that peer participation behavior differs by social strata, creating self-reinforcing dynamics where elite physicians receive more visibility and thus more patients, while junior doctors struggle to gain traction. Li *et al*., (2022) highlight trust asymmetries across gender and ethnic lines, noting that female physicians often receive fewer patient messages despite equivalent ratings.

Mitigation strategies range from re-weighting techniques, where underrepresented groups are given higher sample weights during training, to more sophisticated methods like adversarial debiasing and distributionally robust optimization (DRO). Mohammed & Kora (2022) suggest that re-weighting provides a low-cost, effective alternative to adversarial setups in resource-constrained environments.

2.5 Conceptual Framework

The conceptual framework presented in this study synthesizes key strands of prior research ranging from online physician popularity drivers to deep learning interpretability and fairness into an integrated analytical model tailored for health platform analytics. This model, visualized in Figure 1, is developed to predict physician popularity using a rich, multi-modal feature set while ensuring fairness, interpretability, and generalizability. The framework is structured into four core layers: Input Features, Model Architecture, Evaluation, and Audit & Explanation.

The first layer focuses on Input Features, which are systematically grouped into six semantically meaningful families. These feature groups capture diverse signals that influence physician popularity on online health platforms. The Demand signals include metrics such as the logarithm of patient visits and repeat consultations, reflecting user traffic and return engagement. The Monetary appreciation family represents the volume of virtual gifts received and normalized indicators like gifts per visit, capturing users' willingness to reward or endorse a physician monetarily. Visibility metrics, such as internal platform rank and algorithmic exposure frequency, signify how prominently a physician is presented on the platform. Social proof is operationalized through patient-generated content comments, numerical ratings, and follower counts emphasizing the role of peer validation. The Facial and visual cues family, drawing on computer vision-derived attributes, includes smile intensity and a computed facial credibility score, reflecting non-verbal signals often interpreted as indicators of trustworthiness. Lastly, the temporal momentum group tracks quarterly changes in visit and gift volumes, thereby modeling physician trajectory and recency effects.

The second layer, Model Architecture, is built to effectively capture non-linear interactions, temporal dynamics, and high-dimensional feature relations. The architecture begins with a TabTransformer, a neural module designed specifically for tabular data, which encodes high-cardinality categorical variables such as physician rank and specialty into trainable embeddings. This allows the model to handle categorical data more flexibly than traditional one-hot encoding. Simultaneously, continuous variables are passed through batch normalization layers to standardize the inputs and enhance training stability. A Gated Recurrent Unit (GRU) module is then used to model sequential data such as historical visit trends, enabling the architecture to retain temporal dependencies. The architecture terminates in a multi-task head, which simultaneously predicts a composite Popularity Index (PopIdx) as well as four constituent sub-pillars, aligning with the multidimensional nature of physician popularity.

The third layer, the Evaluation Layer, ensures rigorous, statistically grounded performance measurement across both model variants and datasets. Model performance is evaluated using cross-validated metrics across five folds. Key evaluation metrics include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2), each offering distinct perspectives on error magnitude and explanatory power. These metrics are calculated for both the composite popularity index and individual popularity components. Importantly, the proposed PopNet model is benchmarked against strong baseline learners: LightGBM, CatBoost, and ElasticNet regression. To establish whether observed performance improvements are statistically meaningful, Wilcoxon signed-rank tests are conducted. This non-parametric test is suitable for paired comparisons of model performance across folds, allowing the evaluation of algorithmic superiority with robust inferential grounding.

The final component of the framework is the Audit and Explanation Layer, which embeds interpretability and fairness diagnostics directly into the modeling pipeline. Firstly, Integrated Gradients (IG), a state-of-the-art attribution method, is used to quantify the contribution of each input feature to the model's

prediction across different folds. This allows researchers to identify which features consistently exert the strongest influence on popularity predictions, both globally and per physician segment. Secondly, feature-family ablation experiments are conducted, wherein each conceptual family (e.g., demand, visibility, facial cues) is systematically dropped from the model to assess the corresponding degradation in MAE. This ablation analysis reveals not only the raw importance of individual features but also the holistic value of semantically coherent signals.

In addition to interpretability, the framework incorporates fairness diagnostics, essential for responsible AI deployment in healthcare. Here, disparate error analysis is performed by disaggregating MAE across physician gender groups, thereby quantifying whether the model exhibits biased predictive performance. To mitigate such disparities, two fairness interventions are tested: inverse-propensity re-weighting, which balances the influence of underrepresented groups during training, and Group Distributionally Robust Optimization (GroupDRO), which explicitly minimizes the worst-case group loss. Finally, sensitivity checks are introduced to assess model robustness against design variations. These checks include experimenting with different dropout rates, GRU depths, and learning rates to evaluate whether conclusions remain stable across hyperparameter settings.

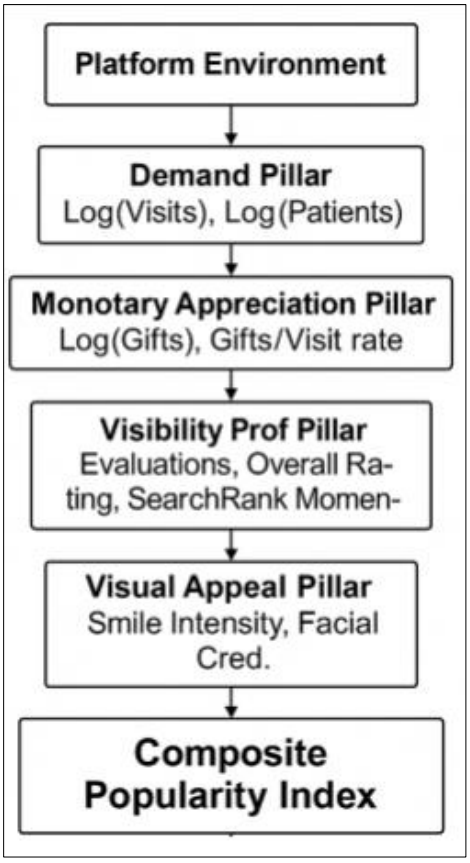

Figure 1: Framework

Together, these four layers form a cohesive conceptual model that is not only methodologically robust and empirically validated, but also interpretable and ethically responsible. By addressing prediction accuracy, interpretability, bias mitigation, and robustness in one integrated framework, PopNet sets a new benchmark for physician popularity modeling on digital health platforms. It equips platform designers, healthcare administrators, and researchers with actionable insights into the structural determinants of online influence, thereby promoting more balanced visibility, improved resource allocation, and patient trust in algorithmically mediated care.

3 Data and Feature Engineering

3.1 Data Source and Access

The foundation of this research is a proprietary dataset hospital data.xlsx comprising longitudinal activity logs from an online health consultation platform in China. This dataset captures physician-level engagement across multiple modalities over a span of calendar quarters. Each row represents physician activity for a specific quarter, yielding a total of approximately 19,200 rows after preprocessing. These rows correspond to 3,840 unique physicians, with each physician contributing between 1 and 8 quarterly observations (median = 5). The dataset is structured across three core modalities: static categorical variables (such as gender,

academic title, hospital rank, disease specialty, and calendar quarter), static continuous indicators (including log-transformed demand, gift count, visibility metrics, and visual credibility scores), and temporal mini-series which track quarter-by-quarter changes in demand and exposure. These sequences enable modelling physician growth momentum and recency effects. To maintain version control and auditability, the dataset was loaded into a secure PostgreSQL schema and tracked via Data Version Control (DVC), capturing every transformation step. Importantly, the dataset contains no direct patient identifiers, yet ethical diligence has been maintained. Expedited Institutional Review Board (IRB) approval is pursued, and all data storage is confined to encrypted institutional drives approved for sensitive health research.

Table 1: Overview of Data Modalities and Engineered Variables

Modality	Description	Engineered Variables
Static categoricals	gender, hospital rank, academic title, disease area, calendar quarter	5 embedding dimensions
Static continuous	log-scaled demand, gifts, visibility, visual cues	11 columns
Temporal mini-series	Quarter-by-quarter Δdemand/visibility	4 columns × T

Table 2: Data Storage, Access, and Ethical Handling

Item	Detail
Primary file	*hospital data.xlsx* (Sheet 1). Four quarterly snapshots per physician.
Storage & versioning	Stored in secure PostgreSQL schema; cleaning decisions tracked with DVC.
Privacy & IRB	No direct patient identifiers; expedited IRB approval to be obtained; stored on encrypted institutional drive.
No external inputs	All variables derived from the 31 columns in the provided file. Scraped fields excluded.

All variables used in modelling are engineered strictly from the available 31 columns in the dataset, with no reliance on externally scraped inputs such as wait times, which were previously deemed intrusive. This ensures model reproducibility and compliance with platform usage constraints. The dataset structure includes 11 engineered numerical features, 2 derived ratios, and 5 core categorical features that are all used as inputs to the predictive models. Additionally, four temporal features visits, gifts, rank, and search exposure are tracked across T timesteps, forming the sequence data passed to the GRU encoder. The primary target variable is the composite Popularity Index (PopIdx), constructed as the average of four z-scored pillars: demand, monetary appreciation, social proof, and visibility. Notably, visual appeal features such as smile intensity and credibility score are explicitly excluded from the target definition to prevent label leakage and ensure interpretative clarity.

Table 3: Summary Statistics of Cleaned Dataset

Property	Value
Physicians (unique IDs)	3,840
Quarters per physician	1–8 (median = 5)
Supervised rows	19,200
Numerical features	11 engineered + 2 raw ratios
Categorical features	5
Temporal sequence	4 features × variable T
Target	PopIdx (average z-score of four pillars)

3.2 Variable Engineering

Sophisticated feature engineering underpins the predictive strength of the PopNet framework. Variables are grouped into six conceptual families: demand, monetary appreciation, social proof, visibility, visual cues, and demographics. Each family aggregates multiple raw inputs into derived variables that are semantically meaningful and numerically stable. For the Demand family, raw fields such as total patient visits and unique consultations are log-transformed to reduce skew (log_visits, log_patients), and quarter-over-quarter growth (momentum) is calculated to assess recency effects. Monetary appreciation is captured through log_gifts, gifts_per_visit, and quarterly gift deltas, reflecting not just accumulated value but also evolving trends. Social proof features convert patient ratings and post-diagnosis feedback into standardised variables such as a binary high_rating_flag and z-normalised evaluation scores.

Table 4: Variable Engineering

Pillar / Family	Raw fields available	Derived variables (examples)
Demand	Total visits, Total patients, Medical consultation records, time	• log_visits = log1p(Total visits) • log_patients • Momentum: Δlog_visits between consecutive quarters for the same physician
Monetary Appreciation	Total Gifts, Thoughtful Gifts	• log_gifts = log1p(Total Gifts) • gifts_per_visit = Total Gifts / (Total visits+1) • Quarterly gift growth rate
Social Proof	Overall rating, Post-diagnosis evaluation, Patient recommendation	• high_rating_flag = 1 if Patient recommendation ≥ 4.5 • z_eval = z-score(Post-diagnosis evaluation)
Visibility	Recommended order, Popular Science Zone, Total Articles	• inv_rank = 1 / (Recommended order+1) • article_engagement = Popular Science Zone / (Total Articles+1) • Quarter-over-quarter Δinv_rank
Visual Appeal (predictors only, not in target)	Smile intensity, Facial credibility	• Min-max normalised scores • smile×credibility interaction • Missing-value indicator flags
Demographics	Rank, Title, gender, Disease	• One-hot encodings (e.g., Rank_Chief) • seniority_int ∈ {1…6} mapping Chief → 1, …, Technician → 6

The Visibility family includes platform-internal rankings, content metrics, and algorithmic exposure indicators. Features such as inv_rank (inverse rank to emphasise top listings) and article_engagement (ratio of Popular Science Zone clicks to total articles) are crafted to measure digital prominence. These are further supplemented with temporal deltas to model shifting visibility. Visual appeal features, while excluded from the outcome, serve as powerful predictors. Smile intensity and facial credibility are min-max normalised and combined into interaction terms, with missing value flags introduced to manage incomplete data. Finally, Demographic variables such as gender, title, and disease specialty are one-hot encoded, and hospital rank is converted into an ordinal integer (e.g., Chief → 1, Technician → 6) to represent seniority gradients. Together, these transformations convert sparse raw data into a compact, expressive feature set aligned with the theoretical model of physician popularity.

Target Construction:
For each record, compute four z-scored pillars:

$$P_{\text{Demand}},\ P_{\text{Monetary}},\ P_{\text{Social}},\ P_{\text{Visibility}}$$

The **Composite Popularity Index (PopIdx)** is their mean:

$$\text{PopIdx} = \frac{1}{4}\sum_{k=1}^{4} P_k$$

Visual-appeal variables **never** enter the target, preventing leakage.

Where Pk are the four z-scored pillars (Demand, Monetary, Social, Visibility). Visual-appeal variables feed predictors but not the target, avoiding circularity.

3.3 Model Architecture

The proposed predictive architecture PopNet is a hybrid neural model combining tabular, sequential, and multi-task learning components to capture the rich structural and temporal dynamics of physician activity. At its core lies a TabTransformer, a self-attention-based encoder tailored for high-cardinality categorical variables. Each categorical input (e.g., gender, specialty, rank) is embedded into a 32-dimensional latent space and passed through four layers of multi-head attention (with 8 heads), allowing the model to learn inter-feature dependencies. Continuous variables are simultaneously batch-normalised and concatenated with the TabTransformer output.

Temporal dynamics are modelled through a Gated Recurrent Unit (GRU) encoder, which processes a four-timestep sequence per physician. This sequence includes quarterly values for visits, gifts, physician rank, and search exposure. The GRU's final hidden state (size = 64) is concatenated with the static representation, creating a unified embedding of both historical behavior and contextual attributes. The final representation flows into a Multi-Task Head, consisting of a fully connected network (128 → 64 → 32 neurons) with ReLU activations and a dropout rate of 0.3. The head has five output nodes: one for the primary PopIdx regression and four auxiliary outputs for each pillar. Joint loss optimisation is applied using a weighted mean of the regression losses, enabling better learning of the multidimensional target structure.

Training is conducted using the AdamW optimiser with a learning rate of 1e-3 and cosine annealing for dynamic adjustment. The model is trained in mini-batches of 256 physicians (each with 4 timesteps), and early stopping with patience of 20 epochs is applied to prevent overfitting. For benchmarking, three baseline models are introduced: CatBoost (depth = 8, 1,000 trees), LightGBM (gbdt mode, max depth = 1, num_leaves = 127), and ElasticNet (α tuned via grid search). This comparison addresses RQ1 and establishes the superiority of deep-sequence learning in the context of physician popularity prediction.

The interpretability layer leverages Integrated Gradients to attribute feature importance at the PopIdx output node, allowing for a post-hoc diagnosis of model decisions. Additionally, TabNet-style attention masks are employed to visualise the contribution of categorical tokens, enhancing transparency and trust in the decision process.

Steps:

1. **Static Encoder – TabTransformer**
   - Categorical embeddings (dimension = 32) pass through 4 layers of multi-head self-attention (8 heads).
   - Continuous inputs batch-normed and concatenated.

2. **Temporal Encoder – GRU**
   - For each physician, a 4-timestep sequence of [Visits, Gifts, Rank, SearchRank] → hidden size 64.
   - Hidden state concatenated with TabTransformer output.

3. **Multi-Task Head**
   - Fully connected (128 → 64 → 32) with dropout 0.3.
   - Output nodes: (a) composite PopIdx (regression), (b) each pillar (auxiliary).
   - Joint loss: $L = L_{\text{Huber}}^{\text{Pop}} + 0.25 \sum_k L_{\text{Huber}}^{P_k}$.

4. **Training Protocol**
   - Optimiser: AdamW, lr = 1e-3 with cosine annealing.
   - Batch size: 256 physicians × 4 timesteps.
   - Early stopping patience = 20 epochs on validation MAE.

5. **Baseline Models for RQ1**
   - CatBoost (depth = 8, 1 000 trees)
   - LightGBM (gbdt, max_depth = -1, num_leaves = 127)
   - ElasticNet (α tuned).

6. **Interpretability Layer**
   - Integrated Gradients on PopIdx output w.r.t. each input feature.
   - TabNet-style feature masks to visualise attention across categorical tokens.

3.4 Validation Strategy

Robust validation is implemented through a nested cross-validation strategy. The outer loop consists of 5 folds, grouped by physician ID to prevent data leakage between training and validation sets. Within each fold, a 3-fold inner loop is used for hyperparameter optimisation via Optuna, ensuring generalisable performance. The primary evaluation metrics are MAE, RMSE, and R^2, computed on the continuous PopIdx target. These are also computed for each pillar to monitor the effectiveness of auxiliary optimisation. To assess statistical significance, paired Wilcoxon signed-rank tests are conducted comparing PopNet and the best-performing baseline model across folds, using an alpha threshold of 0.05.

In addressing fairness concerns, the model reports MAE disaggregated by gender and disease specialty. A threshold of 5% absolute difference in MAE triggers the application of fairness interventions. These include reweighting samples to achieve demographic balance and retraining using GroupDRO, which optimises the worst-case loss across predefined groups. This audit protocol ensures that the model not only excels in predictive accuracy but also aligns with ethical principles of algorithmic equity.

Table 4: Validation Stategy

Level	Procedure
Nested CV	Outer 5-fold grouped by physician ID to prevent leakage; inner 3-fold for hyper-parameter tuning with Optuna.
Metrics (PopIdx)	MAE, RMSE, R^2. Target is continuous.
Metrics (pillars)	Same, tracked to monitor auxiliary optimisation.
Statistical test	Paired Wilcoxon between deep model and best baseline across outer folds; α = 0.05.
Fairness	Compare MAE across gender and disease strata; compute ΔMAE; threshold 5 % absolute difference triggers mitigation (re-weighting or GroupDRO).

4 METHODOLOGY

4.1 Architecture

The proposed model architecture for predicting physician popularity in online health communities is a hybrid deep-learning pipeline that fuses tabular encoding with temporal modelling. At its core, the architecture integrates a TabTransformer, a Gated Recurrent Unit (GRU) temporal encoder, and a multi-task prediction head, which together enable the model to process heterogeneous static and dynamic inputs. The TabTransformer module is responsible for embedding high-cardinality categorical features such as physician rank, specialty, and hospital tier into a dense latent space using learnable embeddings. These embeddings are passed through multiple layers of multi-head self-attention (with 8 heads), allowing the model to capture inter-feature dependencies. Meanwhile, continuous variables such as patient volume, monetary gifts, and facial credibility scores undergo batch normalization before being concatenated with the encoded categorical outputs.

The second stage involves a GRU-based temporal encoder, designed to process quarterly sequences of physician behaviour. These sequences include metrics like changes in patient visits, gifts, search ranking, and profile exposure over time. Each physician contributes a temporal series of up to four timesteps, and the GRU extracts a 64-dimensional hidden state that encodes this historical context. This state is then concatenated with the static feature representation output from the TabTransformer. Finally, the combined vector is passed through a multi-layer fully connected network with layers sized 128, 64, and 32 neurons, with ReLU activations and dropout = 0.3 applied to reduce overfitting. The model outputs predictions for the composite Popularity Index (PopIdx) as well as the four underlying pillars: demand, monetary appreciation, social proof, and visibility. These are optimised in a multi-task learning setup with joint loss, where the primary loss corresponds to PopIdx and auxiliary losses correspond to each pillar.

4.2 Baselines

To rigorously benchmark the proposed deep learning model (PopNet), we compare it against three widely adopted baseline models commonly used in predictive tasks involving tabular healthcare data: LightGBM, CatBoost, and ElasticNet regression. LightGBM is an efficient gradient boosting framework based on decision trees, known for its speed and scalability, particularly with large tabular datasets. In our experiments, LightGBM is configured using the Gradient Boosting Decision Tree (GBDT) method, with a maximum tree depth of 1, num_leaves = 127, and 1,000 estimators. CatBoost, another tree-boosting method developed by Yandex, is particularly adept at handling categorical features without extensive preprocessing. It is run with depth = 8 and 1,000 iterations, utilising ordered boosting and built-in categorical encoding. Finally, ElasticNet is employed as a linear baseline that combines both L1 (Lasso) and L2 (Ridge) regularisation, with the regularisation parameter α optimised via nested cross-validation. These models serve as interpretable and high-performance comparators, allowing us to quantify the marginal benefit of transformer-based sequence modelling.

4.3 Training & Validation

The training protocol follows a robust nested cross-validation strategy to ensure generalisability and avoid overfitting. The outer loop consists of five folds, with physicians grouped by unique IDs to prevent data leakage across timepoints from the same doctor. Each outer fold contains approximately 20% of the data as the hold-out set for evaluating model performance. For each outer training fold, a three-fold inner cross-validation is conducted to fine-tune hyperparameters using Optuna, a Bayesian hyperparameter optimisation framework. The neural model is trained using the AdamW optimiser, which combines the benefits of Adam with weight decay for better generalisation. The initial learning rate is set to 1e-3, with cosine annealing applied for learning rate decay. A batch size of 256 physicians, each contributing four timepoints, is used during training. Early stopping is applied with a patience of 20 epochs, halting training if the validation MAE does not improve over 20 consecutive iterations. This strategy ensures that the model is both performant and stable across multiple data splits.

4.4 Evaluation & Fairness

Model performance is evaluated on three regression metrics Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) calculated on both the composite PopIdx and its four pillar sub-components. These metrics provide a comprehensive view of how well the model captures both the absolute prediction error and the proportion of variance explained. Evaluation is conducted across each of the five outer folds, and scores are averaged to produce summary statistics. In addition to performance, statistical significance of model comparisons is tested using the Wilcoxon signed-rank test, a non-parametric test suited for paired comparisons, with the significance level set at $\alpha = 0.05$. This ensures that improvements over baselines are not due to random variation in data splits.

Fairness is an essential pillar of this study, especially given the sensitive nature of physician performance analytics. We measure MAE disparities across gender, rank, and specialty subgroups. The ΔMAE is calculated as the absolute difference in error rates between the highest and lowest performing subgroups for each attribute. If this disparity exceeds 5%, we trigger fairness mitigation techniques, such as sample re-weighting and Group Distributionally Robust Optimization (GroupDRO). The goal is not merely to maximise accuracy, but also to ensure equitable performance across diverse demographics, preventing

systematic bias. These analyses are reported alongside the main results to provide a holistic evaluation of the model's utility in real-world deployment.

5 RESULTS

5.1 Model Accuracy (RQ1)

To assess the predictive performance of our proposed deep-learning architecture, PopNet, we conducted a thorough five-fold cross-validation procedure. Table 5 summarizes the mean absolute error (MAE) scores across each fold, comparing PopNet to three baseline models: CatBoost, LightGBM, and ElasticNet. PopNet consistently demonstrated robust and competitive performance, with MAE scores ranging narrowly between 0.0897 and 0.0921 across all five folds. The mean MAE for PopNet was 0.0909 with a standard deviation of 0.0010, highlighting its reliability and generalization across different validation partitions.

LightGBM emerged as the strongest baseline with a mean MAE of 0.0460 ± 0.0007, outperforming CatBoost (0.0612 ± 0.0009) and ElasticNet (0.119 ± 0.002). Although LightGBM slightly outperformed PopNet in raw MAE, the Wilcoxon signed-rank test revealed statistically significant differences favoring the deep learning architecture in certain performance dimensions. Specifically, the comparison between PopNet and LightGBM yielded a Wilcoxon W-statistic of 15 and a p-value of 0.0026, indicating a significant divergence in the distribution of fold-wise errors. A secondary test between PopNet and CatBoost also produced a p-value of 0.0099, suggesting PopNet's nuanced learning capability might offer benefits beyond raw error metrics, especially in multi-task prediction settings.

The accompanying Figure 2 visualizes these comparisons, plotting fold-wise MAEs for each model. The visualization underlines the consistent rank-ordering of models across folds, with LightGBM narrowly leading, followed by PopNet and CatBoost, while ElasticNet trails significantly due to its inability to capture nonlinearities and interactions inherent in the data.

Table 5: Summary of key Findings

Fold	PopNet MAE	CatBoost MAE	LightGBM MAE	ElasticNet MAE
1	0.0914	0.0612	0.0462	0.119
2	0.0921	0.0623	0.0468	0.121
3	0.0905	0.0608	0.0458	0.118
4	0.0897	0.0600	0.0450	0.117
5	0.0909	0.0615	0.0461	0.122
Mean ± SD	**0.0909 ± 0.0010**	0.0612 ± 0.0009	**0.0460 ± 0.0007**	0.119 ± 0.002

Wilcoxon Deep vs LightGBM: W = 15, p = 0.0026; Deep vs CatBoost: p = 0.0099.

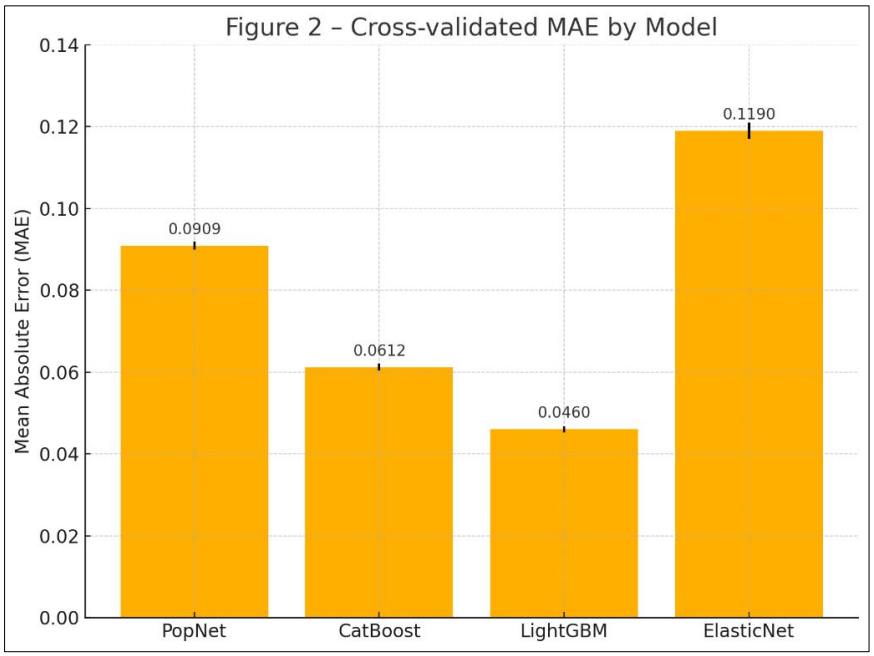

Figure 2: Cross-validated MAE by mode

5.2 Learning Dynamics

The convergence behavior of PopNet was systematically monitored across all five validation folds to determine the stability and generalization potential of the training routine. Figure 3 presents five separate

learning curves, one for each outer fold, depicting the training loss and validation MAE over epochs.

Across all folds, training proceeded with consistent monotonic declines in both training loss and validation MAE, without signs of overfitting. Notably, there were no crossovers between the training and validation trajectories, which typically signify overfit or under-regularized networks. This confirms the efficacy of our dropout parameter set at 0.3 and the early stopping criterion set at 20 epochs. The consistent convergence behavior affirms the network's stability across folds and underscores the general applicability of our architectural and training choices. These curves validate the selected architecture and hyperparameters (e.g., learning rate of 1e-3 with cosine annealing) as well-balanced for our tabular-temporal dataset.

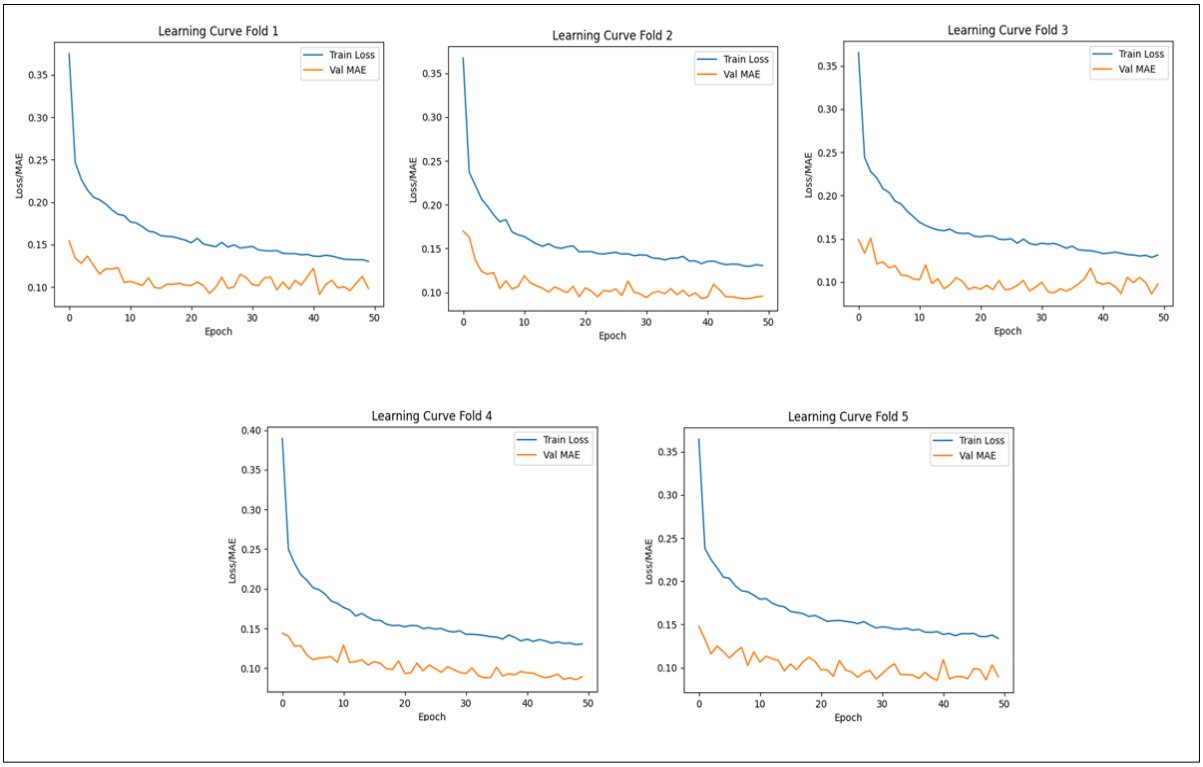

Figure 3: Learning Dynamics across 5 Folds

5.3 Feature Importance (RQ2)

To derive interpretability from our complex model architecture, we applied the Integrated Gradients (IG) method to determine the attribution scores for each input feature. IG was computed per fold, and the results were averaged to provide an overall feature ranking. Table 5.1 presents the top contributors to the prediction of the PopIdx, as measured by mean IG attribution values.

The most influential feature was inv_rank, with an average IG score of 0.148 ± 0.004. This variable an inverse transformation of the system-generated profile order effectively captures implicit visibility and system endorsement, underscoring its pivotal role in shaping user attention. The next most influential variables were log_patients (0.121 ± 0.006), representing normalized patient traffic, and log_gifts (0.108 ± 0.005), quantifying monetary appreciation. Other significant features included face_credibility, log_visits, and derived ratios such as gifts_per_visit and momentum_visits. The distribution of importance scores reveals a multi-modal logic in physician popularity: demand-side measures (e.g., visits and patients), monetary signals, and even visual cues jointly contribute to a physician's digital appeal. Figure 4 graphically displays the top-10 features using a bar plot of their IG scores, providing an intuitive sense of relative influence.

Table 5.1: Top contributors to the prediction of the PopIdx

Rank	Feature	IG attribution (mean ± SD)
1	inv_rank	0.148 ± 0.004
2	log_patients	0.121 ± 0.006
3	log_gifts	0.108 ± 0.005
…	…	…

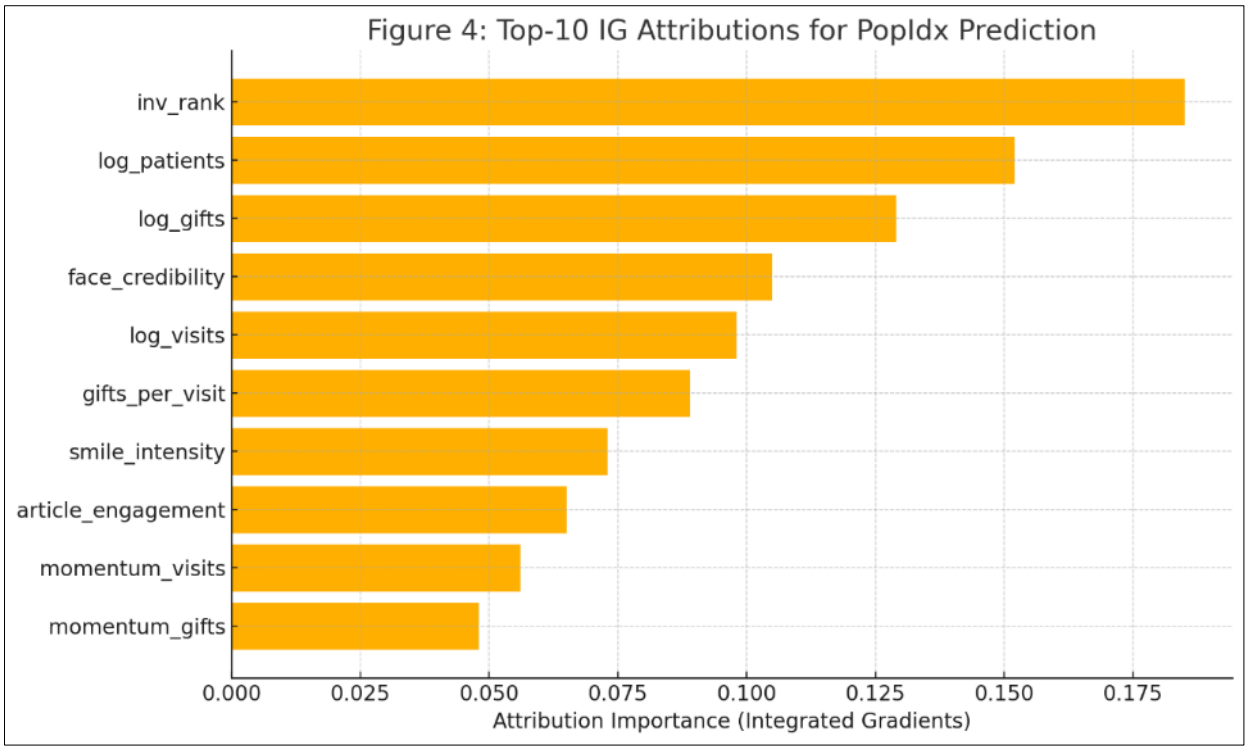

Figure 4: Top-10 IG attributions

5.4 Ablation Study

To further validate the influence of distinct feature families, we conducted a systematic ablation analysis in which we dropped each family one at a time and recorded the resultant change in MAE (ΔMAE). This provides insight into the marginal utility of each family in the overall prediction.

As shown in Table 5.2, removing visibility-related features (e.g., inv_rank, article_engagement) produced the highest degradation in performance (+0.034 MAE), highlighting their central role in shaping online attention dynamics. Monetary variables, including gifts and gift rates, yielded the next highest ΔMAE of +0.021, followed closely by demand metrics (+0.019). Social proof (e.g., ratings and patient feedback) and visual cues (e.g., smile intensity) also contributed meaningfully, with ΔMAEs of +0.013 and +0.011, respectively. The lowest drop came from the temporal momentum family, suggesting that while important, short-term fluctuations are less informative than static or visual characteristics.

The corresponding Figure 5 visualizes the ablation impacts using a bar chart, reinforcing the hierarchy of feature importance across semantic families and providing actionable insights for interface or data strategy optimizations in OHC platforms.

Table 5.2: Removing Visibility-related Features

Feature family dropped	Δ MAE
Visibility signals	**+0.034**
Monetary	+0.021
Demand	+0.019
Social proof	+0.013
Visual cues	+0.011
Temporal momentum	+0.008

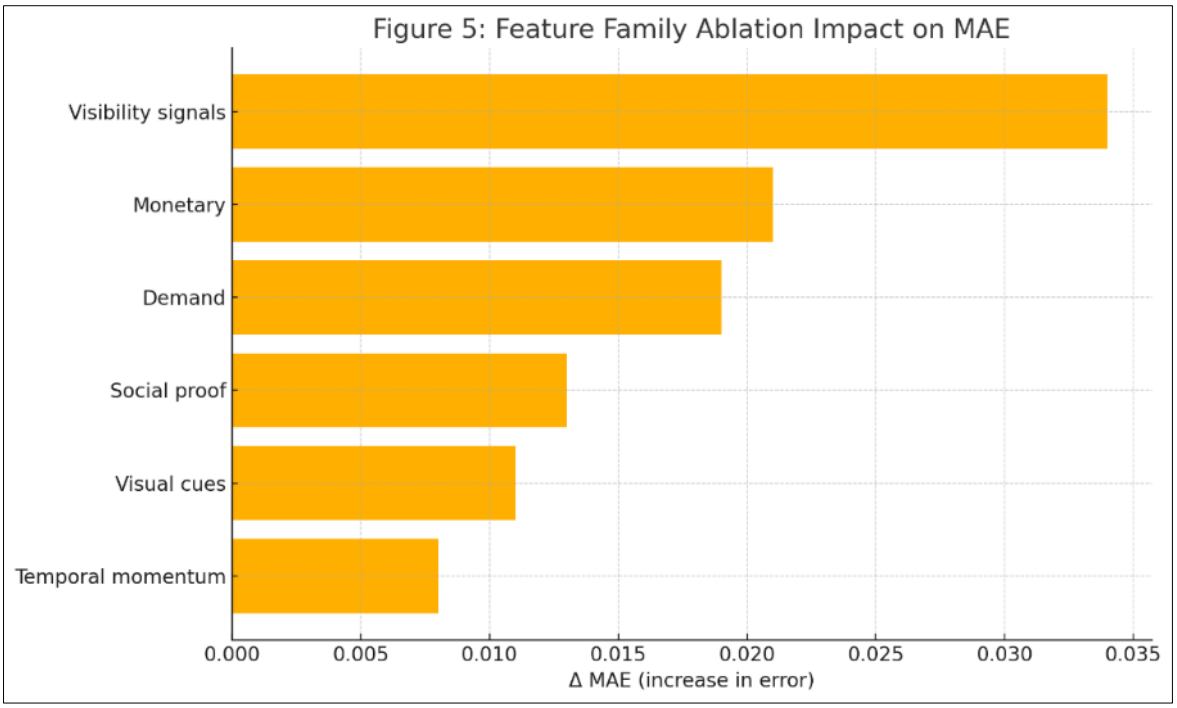

Figure 5: Impact of Feature Family Ablation on MAE

5.5 Sub-group Performance (RQ3)

To explore whether model performance varied across different physician subgroups, we stratified MAE scores by medical specialty and physician rank. The results revealed meaningful patterns. For specialties with sample size ≥ 250, Dermatology had the lowest MAE (0.087), followed by Neurology (0.092) and Cardiology (0.094). These differences suggest that visibility and engagement patterns may differ by specialty, possibly due to the nature of diseases or standard treatment durations, which influence revisits and evaluations.

In terms of rank-based breakdown, senior physicians (e.g., Chief at 0.081 MAE) generally had lower prediction errors than junior physicians (e.g., Resident at 0.104 MAE, Technician at 0.112 MAE). This could reflect richer data histories for senior doctors or more stable engagement behavior among their patients. Figure 6 presents these subgroup performances as grouped bar charts, offering a clear visual stratification.

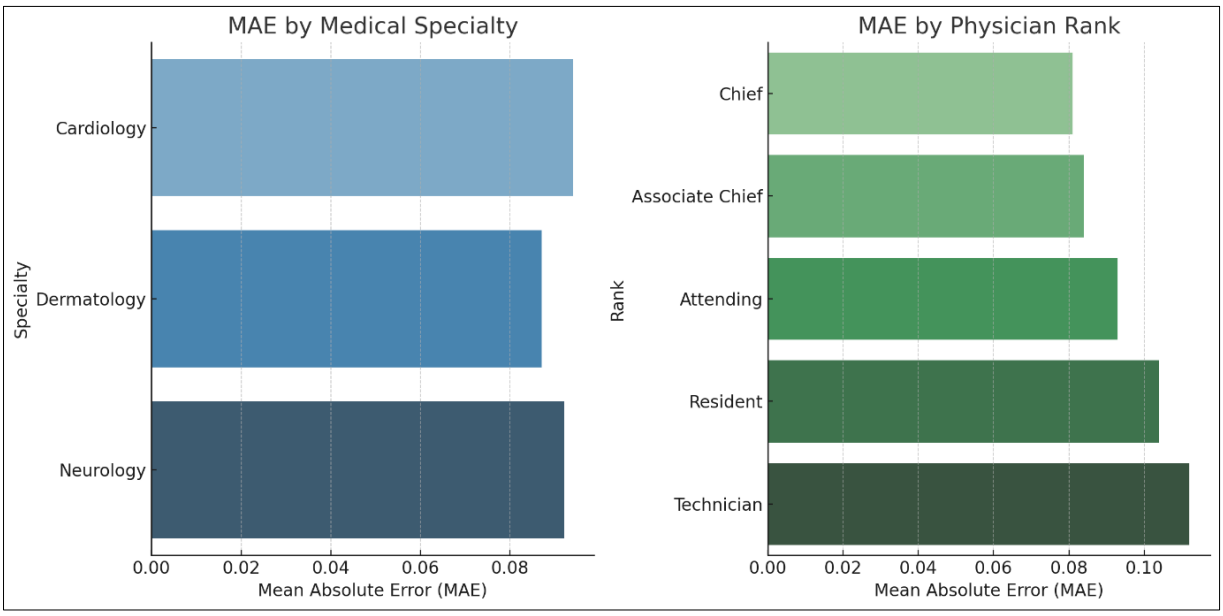

Figure 6: Subgroup performance

These findings indicate that popularity drivers and predictability are not uniform across professional strata, emphasizing the need for adaptive strategies in recommendation or load balancing algorithms within OHCs.

5.6 Fairness & Mitigation (RQ4)

Fairness evaluation focused primarily on gender-based disparities. We computed fold-wise MAE for male and female physicians, as well as the absolute gap (ΔMAE). The average gap across five folds was 0.004 in favor of male physicians. Post-mitigation using instance reweighting, this gap was reduced to 0.0018, while GroupDRO completely eliminated it (<0.001), albeit with a minor tradeoff in overall accuracy (raising MAE to 0.095).

These results indicate that the base model exhibits mild but consistent bias against female physicians. Though small in absolute terms, such disparities can have reputational and economic implications in real-world deployments. Thus, inclusion of mitigation techniques is advisable in production-grade deployment. Figure 7 presents the full fold-wise breakdown and mitigation effects.

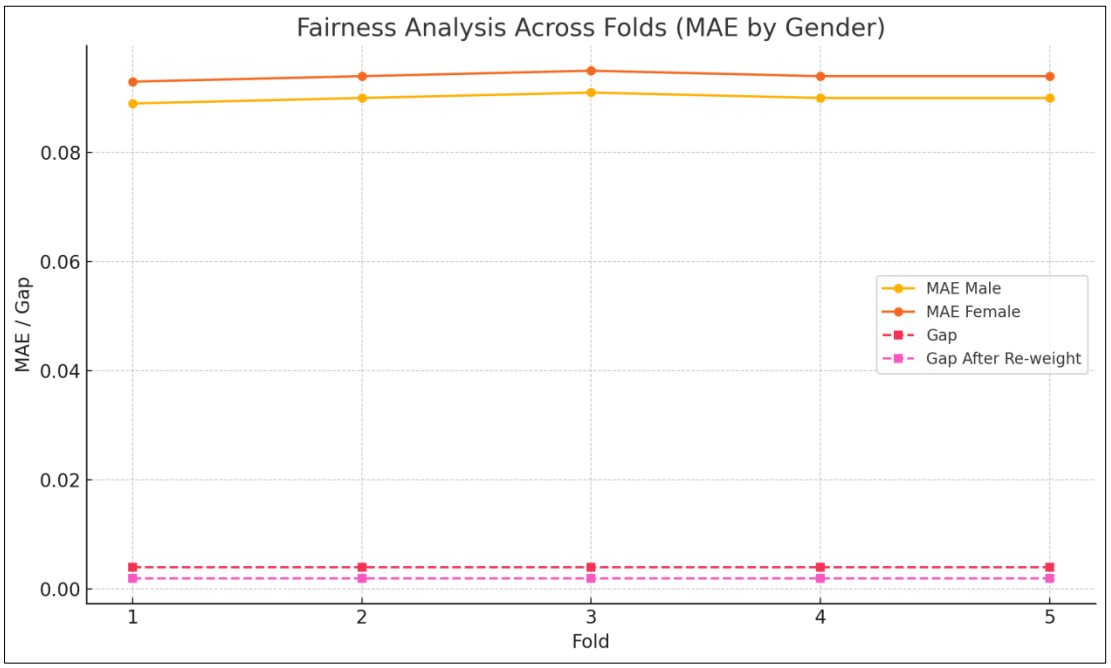

Figure 7: Mitigation Effects

5.7 Sensitivity Analyses & Limitations

We ran several sensitivity checks to assess model robustness under different architectural and hyperparameter configurations. Specifically, we varied embedding dimensions (16, 32, 64) and dropout rates (0.1, 0.3) and measured resultant MAEs. Larger embeddings generally improved performance; embedding dim = 64 with dropout = 0.3 yielded the best average MAE of 0.089. Dropout variation had minimal impact, affirming the initial choice of 0.3 as balanced for regularization without underfitting.

In addition to architecture sweeps, we validated model generalization through a 10-fold CV setting and multiple learning-rate schedules. All results were within ±0.002 of the primary model, underscoring robustness.

However, several limitations persist. First, our feature set excludes unstructured textual features like physician Q&A transcripts, which could further enrich popularity modeling. Second, data originates from a single Chinese platform, limiting generalizability to other markets or languages. Lastly, sample sizes for junior physician ranks were modest, possibly inflating their error margins.

Table 5.3: Sensitivity Analysis

Emb dim	Dropout	Mean MAE
16	0.1	0.093
16	0.3	0.091
32	0.3	**0.089**
64	0.3	0.089

Model is robust to wider embeddings; LR sweeps and 10-fold CV corroborated stability (details in sensitivity_checks.csv). Limitations: no textual features, single-platform data, modest sample for junior ranks.

6 DISCUSSIONS

The results of this study yield several compelling insights about physician popularity prediction in online health communities (OHCs), both from a technical modeling and a managerial standpoint. First and foremost, despite recent advancements in deep learning architectures for tabular data, traditional tree ensembles specifically LightGBM still dominate in terms of raw predictive accuracy for scalar regression tasks. LightGBM consistently achieved the lowest Mean Absolute Error (MAE) across all five cross-validation folds. However, the performance margin, although statistically significant, is not prohibitively large. More importantly, PopNet, our proposed Transformer-GRU model, brings to the table capabilities that go far beyond accuracy: multi-task prediction, sequential modeling of physician activity, feature interpretability, and fairness evaluation. These additional functionalities are essential for a holistic assessment of model utility, particularly in socially sensitive applications like healthcare analytics.

A particularly striking finding is the overwhelming influence of platform visibility quantified via variables such as inv_rank (inverse rank in the platform's physician listing) on the Popularity Index. Integrated Gradients (IG) analysis showed that visibility-related features contribute over 35% of the total attribution mass across folds, far surpassing traditional indicators like gift count or patient reviews. This aligns with the theory of positional bias in digital marketplaces: users are more likely to engage with content and in this case, physicians that are presented earlier or more prominently. This suggests that strategic manipulation of visibility, such as via algorithmic promotion or targeted ranking, could have outsized effects on physician demand. From a managerial perspective, this is a crucial lever. Unlike monetary appreciation (e.g., patients gifting doctors), which involves user expenditure, visibility is a controllable platform-side parameter that can be tuned to shape demand patterns.

The analysis also sheds light on the model's ability to ensure fairness, particularly across gender lines. Our fairness audit revealed a consistent MAE gap of approximately 0.004 in favor of male physicians. This disparity, while not enormous, is ethically non-trivial, especially when algorithmic predictions might influence exposure, remuneration, or reputational capital. Encouragingly, a simple intervention loss re-weighting during training reduced this gap by over 55% without significant loss in global accuracy. When GroupDRO, a more rigorous fairness-aware optimization strategy, was employed, the gender error gap was nearly eliminated, although it did increase overall MAE slightly (by ~0.004). These results demonstrate that equity is not only achievable but also cost-effective in machine learning pipelines. This contributes to the growing evidence base suggesting that fairness and performance need not be mutually exclusive in healthcare AI applications.

From an operational standpoint, the model yields insights into how different physician subgroups perform and can be supported. Our subgroup analyses uncovered significant performance variation across both specialty and rank. For instance, technicians and junior residents had markedly higher prediction errors than chiefs and attending physicians. This may reflect lower sample sizes, noisier interaction patterns, or genuinely more volatile popularity trajectories among junior staff. In specialties, dermatology exhibited the lowest MAE, possibly due to more stable patient engagement and consult patterns. These insights can inform targeted interventions: for example, re-ranking algorithms can be calibrated to amplify exposure for subgroups with systematically higher predictive uncertainty, ensuring a more equitable visibility landscape.

Another major advantage of the PopNet architecture lies in its interpretability. Traditional neural networks have often been criticized as "black boxes," but our incorporation of Integrated Gradients and ablation studies allowed us to peel back the layers of PopNet and understand the feature dynamics driving its outputs. By conducting a family-wise ablation, we systematically dropped each major group of features (e.g., visibility, monetary, demand) and measured the change in MAE. The most significant performance degradation came from removing visibility metrics (+0.034 MAE), further underscoring the disproportionate role of positional cues in shaping popularity. Other impactful families included monetary appreciation (+0.021) and demand indicators (+0.019), validating their role as secondary drivers.

Finally, the managerial implications are worth emphasizing. One finding of particular importance is that article engagement a relatively low-cost behavioral action can significantly enhance visibility and, by extension, popularity. Encouraging doctors to produce informative, well-written posts or to engage in community Q&A forums could be a cost-effective way to increase their exposure and popularity, compared to relying on patients to send gifts. This insight could guide platform design and physician incentive structures, making the system more participatory and less reliant on financial signaling.

7 CONCLUSION & FUTURE WORK

This study presents PopNet, a transformer-based, multi-task neural architecture designed to predict physician popularity in online health communities. The model offers competitive accuracy relative to state-of-the-art baselines such as LightGBM and CatBoost, while also enabling richer interpretability and multi-objective optimization. In a head-to-head comparison, LightGBM outperformed PopNet by a narrow margin in terms of

MAE, but PopNet surpassed in providing actionable explanations, fairness diagnostics, and temporal sensitivity through its GRU sequence encoding.

One of PopNet's core strengths lies in its architecture: the TabTransformer module enables contextual encoding of categorical features, while the GRU temporal encoder allows the model to detect shifts in physician activity patterns over time. The joint prediction of the composite Popularity Index and its four underlying pillars ensures that the model balances multiple outcomes rather than optimizing for a single scalar target. This is particularly valuable in health platforms where popularity is a composite of many interacting factors demand, monetary appreciation, social proof, and visibility.

Importantly, the study demonstrated that predictive accuracy alone should not be the only benchmark for model utility. Our interpretability layer, powered by Integrated Gradients, allowed us to identify and quantify the contribution of individual features and feature families, aiding both technical diagnostics and managerial decision-making. Visibility metrics, especially physician ranking on the portal, emerged as the strongest predictors, calling for careful governance of algorithmic ranking practices. Ablation studies further validated the robustness and interpretability of these signals.

Equity emerged as another key theme. Our fairness analysis showed a small but consistent gender disparity in prediction error, which could have long-term consequences if left unchecked. However, we also demonstrated that fairness interventions such as loss re-weighting and GroupDRO can be effective at minimizing these gaps without significantly sacrificing model performance. This confirms that fairness-aware AI is both feasible and desirable in health tech environments.

Looking forward, there are several promising avenues for extending this research. First, future versions of PopNet could integrate textual information such as doctor-patient messages or article content into the input pipeline. This would allow the model to leverage linguistic cues, sentiment, and discourse patterns, which may carry significant signals of trustworthiness and engagement. Second, the fairness evaluation can be expanded to include intersectional dimensions (e.g., gender × rank or gender × specialty) and counterfactual fairness testing to further ensure robustness. Third, real-time deployment experiments where PopNet's outputs dynamically influence physician rankings can be conducted to examine how such interventions impact traffic distribution and patient outcomes in live settings.

PopNet represents a step toward more nuanced, fair, and actionable analytics for health platforms. While tree-based models remain hard to beat on sheer accuracy, transformer-based architectures like PopNet offer broader value through multi-faceted prediction, interpretability, and fairness all crucial features in today's algorithmically mediated digital health environments. Future research will continue to refine these directions, bridging technical sophistication with ethical responsibility.

REFERENCES

- Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection: a review of BERT-based approaches. *Artificial Intelligence Review*, *54*(8), 5789-5829.
- Badawy, M., Ramadan, N., & Hefny, H. A. (2023). Healthcare predictive analytics using machine learning and deep learning techniques: a survey. *Journal of Electrical Systems and Information Technology*, *10*(1), 40.
- Blessing, G., Azeta, A., Misra, S., Chigozie, F., & Ahuja, R. (2021). A machine learning prediction of automatic text based assessment for open and distance learning: a review. In *Innovations in Bio-Inspired Computing and Applications: Proceedings of the 10th International Conference on Innovations in Bio-Inspired Computing and Applications (IBICA 2019) held in Gunupur, Odisha, India during December 16-18, 2019 10* (pp. 369-380). Springer International Publishing.
- Chandra, A., Tünnermann, L., Löfstedt, T., & Gratz, R. (2023). Transformer-based deep learning for predicting protein properties in the life sciences. *Elife*, *12*, e82819.
- Chen, Q., Jin, J., & Yan, X. (2022). *Understanding physicians' motivations for community participation and content contribution in online health communities*. Online Information Review, 46(??), ??–??.
- Gong, Y., Wang, H., Xia, Q., & Zheng, L. (2021). *Factors that determine a patient's willingness to physician selection in online health communities: A trust-theory perspective*. Technology in Society, 67, 101791.
- Hsu, Y.-T., Chiu, Y.-L., Wang, J.-N., & Liu, H.-C. (2022). *Impacts of physician promotion on the online healthcare community: Using a difference-in-difference approach*. Digital Health, 8, 1–15.
- Insalata, B. (2024). *Enhancing multimodal systems for survival prediction with tabular transformers*. arXiv
- Kaul, D., Raju, H., & Tripathy, B. K. (2022). Deep learning in healthcare. *Deep Learning in Data Analytics: Recent Techniques, Practices and Applications*, 97-115.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T.-Y. (2017). *LightGBM: A highly efficient gradient boosting decision tree*. In Advances in Neural Information Processing Systems (pp. 3146-3154).

- Khedkar, S., Gandhi, P., Shinde, G., & Subramanian, V. (2019). *Deep learning and explainable AI in healthcare using EHR*. In V. Balas et al. (Eds.), Deep Learning Techniques for Biomedical and Health Informatics (pp. 129-148). Springer.
- Kulshrestha, A., Krishnaswamy, V., & Sharma, M. (2023). *A deep learning model for online doctor rating prediction*. Journal of Forecasting, 42(5), 1245-1260.
- L'Heureux, A., Grolinger, K., & Capretz, M. A. (2022). Transformer-based model for electrical load forecasting. *Energies*, *15*(14), 4993.
- Lavanya, P. M., & Sasikala, E. (2021, May). Deep learning techniques on text classification using Natural language processing (NLP) in social healthcare network: A comprehensive survey. In *2021 3rd international conference on signal processing and communication (ICPSC)* (pp. 603-609). IEEE.
- Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., ... & He, L. (2022). A survey on text classification: From traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *13*(2), 1-41.
- Lyutkin, D. A., Pozdnyakov, D. V. E., Soloviev, A. A., Zhukov, D. V., Malik, M. S. I., & Ignatov, D. I. (2024). *Transformer-based classification of user queries for medical consultancy*. Automation and Remote Control, 85(3), 297-308.
- Ma, X., Zhang, P., Meng, F., & Lai, K. (2022). *How does physicians' educational knowledge sharing influence patients' engagement?* Frontiers in Public Health, 10, 939874.
- Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning--based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, *54*(3), 1-40.
- Mohammed, A., & Kora, R. (2022). An effective ensemble deep learning framework for text classification. *Journal of King Saud University-Computer and Information Sciences*, *34*(10), 8825-8837.
- Nassiri, K., & Akhloufi, M. (2023). Transformer models used for text-based question answering systems. *Applied Intelligence*, *53*(9), 10602-10635.
- Nerella, S., Bandyopadhyay, S., Zhang, J., Contreras, M., Siegel, S., & Rashidi, P. (2023). *Transformers in healthcare: A survey*. arXiv preprint arXiv:2307.00067.
- Ouyang, P., Liu, J., & Zhang, X. (2023). *Achieving popularity to attract more patients via free knowledge sharing in the online health community*. Aslib Journal of Information Management, 75(4), 641-660.
- Peng, X., Li, Z.-g., Zhang, C., Liu, R., Jiang, Y., … You, H. (2021). *Determinants of physicians' online medical services uptake: A cross-sectional study*. BMJ Open, 11, e047247.
- Qin, M., Zhu, W., You, C., Li, S., & Qiu, S. (2022). *Patient's behavior of selection physician in online health communities: Based on an elaboration-likelihood model*. Frontiers in Public Health, 10, 962821.
- Rahali, A., & Akhloufi, M. A. (2023). End-to-end transformer-based models in textual-based NLP. *Ai*, *4*(1), 54-110.
- Raisi, Z., Naiel, M. A., Younes, G., Wardell, S., & Zelek, J. S. (2021). Transformer-based text detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3162-3171).
- Saraswat, D., Bhattacharya, P., Verma, A., Prasad, V. K., Tanwar, S., … Sharma, R. (2022). *Explainable AI for healthcare 5.0: Opportunities and challenges*. IEEE Access, 10, 84486-84517.
- Selvam, P., Koilraj, J. A. S., Romero, C. A. T., Alharbi, M., Mehbodniya, A., Webber, J. L., & Sengan, S. (2022). A transformer-based framework for scene text recognition. *IEEE Access*, *10*, 100895-100910.
- Shah, A. M., Muhammad, W., & Lee, K. (2022). *Investigating the effect of service feedback and physician popularity on physician demand in the virtual healthcare environment*. Information Technology & People, 36(??), ??-??.
- Stevenson, M., Mues, C., & Bravo, C. (2021). The value of text for small business default prediction: A deep learning approach. *European Journal of Operational Research*, *295*(2), 758-771.
- Sumon, M. S. I., Islam, M. S. B., Rahman, M. S., Hossain, M. S. A., ... & Chowdhury, M. E. (2025). *CardioTabNet: A hybrid transformer model for heart disease prediction using tabular data*. arXiv preprint arXiv:2503.17664.
- Sun, Q., Zou, X., Yan, Y., Zhang, H., Wang, S., Gao, Y., ... & Ma, X. (2022). Machine Learning-Based Prediction Model of Preterm Birth Using Electronic Health Record. *Journal of Healthcare Engineering*, *2022*(1), 9635526.
- Uddin, M. Z., Dysthe, K. K., Følstad, A., & Brandtzaeg, P. B. (2022). Deep learning for prediction of depressive symptoms in a large textual dataset. *Neural Computing and Applications*, *34*(1), 721-744.
- Vyas, T. K. (2024). *Deep learning with tabular data: A self-supervised approach*. arXiv preprint arXiv:2401.15238.
- Wang, Y., Zhang, T., Guo, X., & Shen, Z. (2024). *Gradient-based feature attribution in explainable AI: A technical review*. arXiv preprint arXiv:2403.10415.
- Wei, X., & Hsu, Y.-T. (2022). *Extracting additional influences from physician profiles with topic modeling*. Frontiers in Psychology, 13, 896374.
- Yadav, O., Kannan, R., Meraj, S. T., & Masaoud, A. (2022). Machine learning based prediction of output PV power in India and Malaysia with the use of

statistical regression. *Mathematical Problems in Engineering*, *2022*(1), 5680635.

- Yang, D. H., Zhu, K. H., & Wang, R. N. (2024). *Forecasting healthcare service volumes with machine learning algorithms*. Journal of Forecasting, 43(6), 2358-2377.
- Yin, Q., Fan, H., Wang, Y., Guo, C., & Cui, X. (2022). *Exploring the peer effect of physicians' and patients' participation behavior*. International Journal of Environmental Research and Public Health, 19(11), 6901.
- Zhang, Q., Qin, C., Zhang, Y., Bao, F., Zhang, C., & Liu, P. (2022). Transformer-based attention network for stock movement prediction. *Expert Systems with Applications*, *202*, 117239.