

PAPER

Detecting Health Misinformation Using Retrieval-Augmented Large Language Model

Masuma Mammadova ,
Vagif Mammadaliyev  (✉),
Vusal Shahbazov 

Azerbaijan National Academy
of Sciences, Baku, Azerbaijan

v.mammadaliyev@iit.science.az

ABSTRACT

Misinformation is common in online media, including in health-related content. Such misinformation constitutes a significant risk to the medical-demographic security of the population. Many people actively seek medical information online without the ability to verify its trustworthiness. This paper proposes a modular system for automated detection of misinformation in medical content based on retrieval-augmented verification. In the pipeline, input articles are summarized and decomposed into atomic claims. Each claim is then vectorized using dense embeddings (BGE-Large-en-v1.5) and matched against a self-constructed knowledge base via cosine similarity to retrieve evidence. Retrieved evidence is verified using large language model (LLM)-based natural language inference (NLI) and aggregated into article-level predictions through a scoring function. The LLM used in the system is the open-source LLaMA-3.1 model with eight billion parameters. Together with a self-contained knowledge base, this represents a deliberately constrained environment. Such a setup isolates the contribution of the presented architecture from its dependency on external knowledge sources and LLMs with large parameter counts. The system achieved 98.4% accuracy and a macro F1 of 0.978 at 79.0% coverage on the evaluated datasets (DETERRENT and CONSTRAINT). This demonstrates that, even under resource constraints, the correct pipeline architecture enables high-precision misinformation detection on medical content. It also represents a clear improvement over the ablation experiment, which used the same model without the proposed pipeline and achieved an F1 of only 0.773.

KEYWORDS

medical-demographic security, health misinformation detection, retrieval-augmented verification, claim decomposition, natural language inference (NLI), large language models (LLM), knowledge base construction

1 INTRODUCTION

Medical-demographic security is defined as a state of the population, which is protected from various threats affecting its structure, size and quality. Misinformation in the medical domain, being a significant threat, requires an effective detection

Mammadova, M., Mammadaliyev, V., Shahbazov, V. (2026). Detecting Health Misinformation Using Retrieval-Augmented Large Language Model. *International Journal of Online and Biomedical Engineering (iJOE)*, 22(10), pp. 109–126. <https://doi.org/10.3991/ijoe.v22i10.61955>

Article submitted 2026-04-14. Revision uploaded 2026-07-07. Final acceptance 2026-07-07.

© 2026 by the authors of this article. Published under CC-BY.

mechanism to protect the population from its harmful consequences [1]. The spread of health misinformation during disease outbreaks is characterized as an *infodemic* by the World Health Organization, which also notes that such information affects health behaviours at scale [2]. For instance, during the COVID-19 (Coronavirus disease of 2019) pandemic, 51% of vaccine-related posts and 28.8% of general COVID-19 content on social media were classified as misinformation [3]. Such exposure to vaccine-related misinformation reduced the population's intent to vaccinate by 6.2 percentage points in controlled experiments [4]. Although the current paper focuses on text-based content, the problem of misinformation extends to other media formats. A systematic review of medical content on the YouTube platform found that 41% of assessed videos contained misleading information [5].

The apparent scale of this problem demands detection systems, which are able to identify misinformation with high precision. Traditional approaches to this task rely on supervised classifiers trained on textual features, leveraging pre-trained language models such as BERT and RoBERTa [6], [7]. Although these approaches are able to demonstrate decent results, they lack the capacity to *explain* why specific medical content was classified as misinformation. This happens because they operate on learned statistical patterns without verifying factual claims. Alternative methods rely on other social-context signals, such as user profiles, network structure, and the pattern of content propagation to improve accuracy of detection. However, such features are not always available in real-world settings and are highly dependent on platform-specific metadata [1].

Recent approaches explore LLMs as detectors of misinformation. A particularly promising direction involves the use of retrieval-augmented generation (RAG), in which external evidence is retrieved at inference time [8], [9], [10]. This mechanism has been applied to fact-checking in the medical domain, where such retrieval-based verification reduces large language models (LLM) hallucination and improves factual reliability [11], [12]. The idea of processing article content at the claim level enables more granular and interpretable assessments through natural language inference [13].

However, several challenges regarding RAG-based health misinformation detection systems remain unaddressed. First, existing systems often rely on external knowledge bases, which introduces a significant dependency. Second, there is a tendency in LLMs to soften the language of processed content, which can obscure signals needed for misinformation detection [14]. Third, few systems integrate all parts of the pipeline, such as claim extraction, evidence retrieval, and article-level verdict determination, in a reproducible framework. As a result, it is difficult to assess the contribution of individual components. The problem addressed in this work is the detection of health misinformation from text content alone, without any dependency on online external knowledge sources or on resource-intensive LLMs. Given this problem, the central research question is whether a modular claim-level verification pipeline can achieve competitive accuracy using only an 8B-parameter open-source model and a self-constructed knowledge base.

This paper proposes a modular retrieval-augmented verification system for health misinformation detection, to address this question. The first process in the system is offline knowledge base construction, in which articles taken from datasets are summarized, decomposed into atomic claims, and vectorized into a dense embedding store. The second process is the article verification itself. The article verification process also includes summarization and claim decomposition steps from the first process. The resulting claims are then matched against the constructed knowledge base via semantic similarity search and verified through LLM-based natural

language inference (NLI). The verified claims are aggregated into an article-level prediction through a weighted scoring function.

The system achieves 98.4% accuracy and a macro F1 score of 0.978 at 79.0% coverage across both evaluation datasets, DETERRENT [6] (cancer and diabetes health news articles) and CONSTRAINT [7] (COVID-19 social media posts). These results demonstrate that even under significant resource constraints, the pipeline design achieves competitive accuracy.

The remainder of this paper is organized as follows: Section 2 reviews related work on LLM-based and retrieval-augmented approaches to misinformation detection. Section 3 describes the datasets, the system architecture, and the weighted scoring mechanism. Section 4 presents the experimental results and an ablation study comparing the proposed pipeline against a zero-shot baseline. Section 5 discusses the main findings. Section 6 addresses limitations and proposes future directions. Section 7 concludes the paper.

2 RELATED WORKS

Chen and Shu highlight key challenges in architectures based on LLMs, such as hallucination, sycophantic behaviour, and the difficulty of keeping up to date knowledge. They also characterize LLMs as being capable of both detecting false content and generating it [8]. In contrast to traditional approaches such as content based classifiers, a growing body of studies has responded to these challenges by basing detection decisions on explicitly retrieved evidence. Since its introduction, RAG has been adopted for fact checking and misinformation detection [10]. A comprehensive survey of such approaches is provided by Gao et al. [15]. The survey identifies them as a general solution to overcome LLM hallucination and knowledge staleness.

Several systems retrieve evidence from different external sources to support misinformation detection. Hang et al. introduced TrumorGPT, a framework that constructs semantic knowledge graphs from DBpedia and verifies claims via subgraph matching using GPT-4 [16]. This system achieved 88.5% accuracy on PolitiFact health statements, although its reliance on an LLM with a relatively large parameter count introduces a strong dependency. Bai and Fu instead extract triples from articles in their FCRV framework [17]. They retrieve evidence using live web search and classify articles through LLM reasoning over the full context. FCRV reported an accuracy of 88.0% on GossipCop, but its reliance on web search introduces a different dependency. Retrieved content is not always reliable and may require strong reasoning capabilities from the underlying LLM. Upadhyay and Viviani instead draw evidence from PubMed Central [18]. They combine BM25 retrieval with LLM generated reference texts and rank documents by topical relevance and factual accuracy via stance detection using SciFive. A similar system was introduced by Li et al. for COVID-19 fact checking. Their system draws evidence from a large corpus of peer reviewed papers [11]. Compared with a standalone GPT-4 baseline, their models enhanced with RAG showed substantially higher accuracy and reduced hallucination. Across these systems, the source of evidence varies from knowledge graphs to live web search to curated literature corpora. Each nonetheless introduces a dependency on an external source that must be maintained and trusted.

Beyond the source of evidence, a different line of work examines how claims are decomposed and verified once evidence has been retrieved. Min et al. proposed FActScore, which applies the approach of decomposing text and then verifying it to misinformation detection through NLI [13]. Text generated by an LLM is decomposed into atomic facts, and each fact is then verified against a knowledge source.

Each pairing of claim and evidence is classified as supporting, refuting, or neutral. The introduction of the neutral outcome is particularly important, since it allows the architecture to avoid a binary decision when the evidence is insufficient. Hossain et al. applied the same classification approach to misinformation related to COVID-19 [19]. They constructed a knowledge base of 86 manually curated misconceptions and used NLI models to classify tweets as either agree, disagree, or no stance. While this approach is conceptually aligned with the modular pipeline proposed in the current paper, COVIDLies achieved only 50.2% macro F1 on the stance detection task. Performance on the classes other than “no stance” is particularly poor, with 41.2% F1 for agree and 20.3% for disagree. These are the classes most relevant to misinformation detection. Separately, Ma et al. demonstrated that fine-grained evidence retrieval at the sentence level improves verification accuracy compared with approaches that operate at the document level [20]. This finding is also aligned with the system proposed in the current paper. Their proposed system used hierarchical attention networks to embed evidence at the sentence level for claim verification. Together with FActScore’s atomic decomposition, this suggests that verifying evidence at a finer granularity than the whole document improves reliability.

Beyond individual systems, broader evaluations of LLMs as fact checkers point to similar concerns. Inconsistent accuracy across domains and a tendency of LLMs to classify claims that sound factual as true are pointed out by Quelle and Bovet in their systematic evaluation [12].

The most closely related approaches are summarized in Table 1. Differences in architecture and resource requirements mean this is not a strict point by point comparison. The system proposed in the current paper builds on these findings by avoiding dependence on an external knowledge base or a large LLM. It further differs in its claim decomposition strategy and verification method.

Table 1. Comparison of the most closely related systems

System	LLM	Knowledge Base	Claim Decomposition	Verification Method	Best Accuracy
DETERRENT [6]	–	Medical KG	No	Graph attention	N/A (F1-based)
COVIDLies [19]	–	86 misconceptions	No	NLI stance detection	50.2% macro F1
TrumorGPT [16]	GPT-4	DBpedia	No	GraphRAG matching	88.5%
FCRV [17]	LLaMA-3.1	Live web search	Triples	RAG + LLM reasoning	88.0%
Li et al. [11]	LLM + RAG	Peer-reviewed journals	No	RAG fact-checking	N/A (usability study)
Upadhyay & Viviani [18]	LLaMA-3/GPT-3	PubMed Central	No	Stance + similarity	N/A (IR metrics)
Proposed system	LLaMA-3.1-8B	Self-contained (dataset-derived)	Yes (atomic claims)	Retrieval + NLI + plausibility	98.4%

3 MATERIALS AND METHODS

3.1 Datasets

In this study, two publicly available datasets were used. Both datasets were pre-processed, specifically duplicate entries and articles containing HTTP 404 errors instead of the actual content were removed. Each dataset targets a different format of content.

The DETERRENT dataset [6] used in this study contains 8,132 health news articles related to cancer and diabetes after pre-processing. The original sources of these articles vary; for instance, the dataset contains articles from reliable sources such as WebMD and NIH, and unreliable sources such as NaturalNews and BeforeItsNews. Articles in this dataset are long, a characteristic that plays a significant role in the claim extraction process, as longer articles tend to yield more verifiable claims.

Contrary to DETERRENT, the CONSTRAINT dataset [7] used in this study contains 2,140 short social media posts related to COVID-19. They were collected from online platforms and fact-checking websites. Although the posts are short (approximately 180 characters), limiting the number of extractable claims as discussed in subsequent sections, the label distribution is more balanced compared to DETERRENT.

Statistics for the final evaluation datasets are presented in Table 2.

In the original DETERRENT dataset, cancer- and diabetes-related entries are separated from each other. For the purpose of this study, they were merged together, and the structure of each entry was standardized to contain the following fields: title, content, and reference label. The same standardization rule applies to the CONSTRAINT dataset. To present a clear content format difference between the two datasets, content length distributions are shown in Figure 1.

Table 2. Dataset statistics

Dataset	Articles	Real (Label 0)	Misinformation (Label 1)	Avg. Length (Chars)
DETERRENT	8,132	6,111	2,021	2,843
CONSTRAINT	2,140	1,120	1,020	180
Total	10,272	7,231	3,041	—

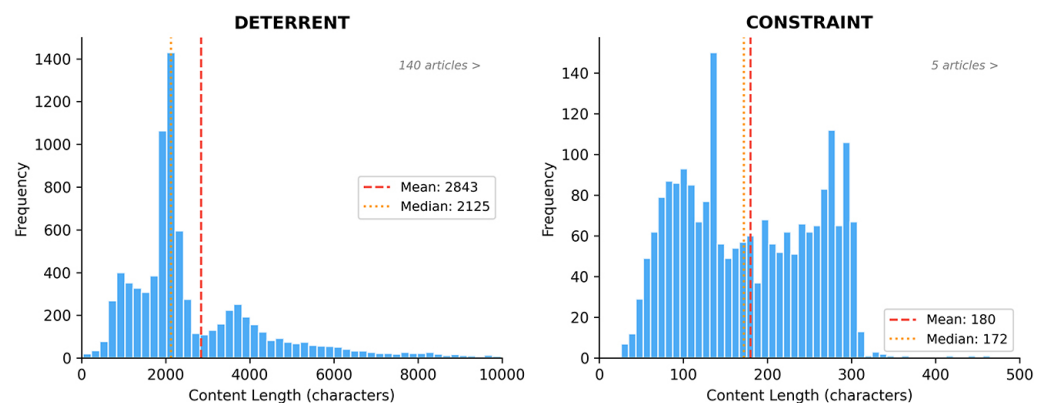


Fig. 1. Content length distributions for DETERRENT and CONSTRAINT

3.2 Pipeline overview

With both datasets in place, the proposed system operates in two main stages: knowledge base construction and article verification. For both stages, summarization and claim extraction logic is shared. LLaMA-3.1-8B performs all LLM-related

tasks, while BGE-Large-en-v1.5 handles all embedding operations. LLM calls performed throughout use a temperature of 0.1 to ensure deterministic, reproducible output.

Knowledge base construction. (see Figure 2) is an offline process. Articles from aforementioned datasets are summarized, preserving only content related to the medical domain, and the resulting summaries are further decomposed into verifiable medical claims. Extracted claims are vectorized and stored along with their reference label. This is a one-time process to construct the knowledge base.



Fig. 2. Knowledge base construction process

In the article **verification process** (see Figure 3), the resulting claims are not stored in the knowledge base but are instead verified by the LLM through NLI using fetched evidence. Finally, claim-level results are processed by a weighted scoring function to produce an article-level classification.

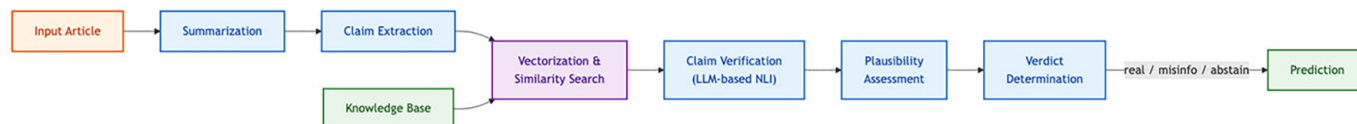


Fig. 3. Article verification process

The pipeline operates only on the self-constructed knowledge base and the content of the articles, without using other signals, such as social-context features and user metadata.

3.3 Article summarization

This section describes the first step shared by both processes: article summarization. Given an article's title and content, the LLM used in this study is prompted to produce a summary, retaining only information related to the medical domain, such as treatments and drug effects, while removing non-medical content, such as opinions and political commentary.

There are two critical design decisions in the summarization prompt. The first is instructing the model to preserve the exact wording of medical claims. For instance, if an article states "X cures cancer," the summary must retain this exact phrasing rather than softening it to "X may have anti-cancer properties." This instruction addresses the tendency of LLMs to soften extreme claims, which is particularly problematic in the misinformation detection context. The second critical design decision is instructing the model to extract only content related to the medical domain. Otherwise, the model returns a special indicator (NO_HEALTH_CONTENT) if there is absolutely no medical information. This design choice supports the abstention principle discussed in Section 5.

As shown in Figure 4, the summarization step identifies 291 articles (3.6%) in DETERRENT and 1,631 posts (76.2%) in CONSTRAINT as containing no extractable medical content. The high number of entries with the NO_HEALTH_CONTENT indicator in the CONSTRAINT dataset is due to the relatively short entries in this dataset.

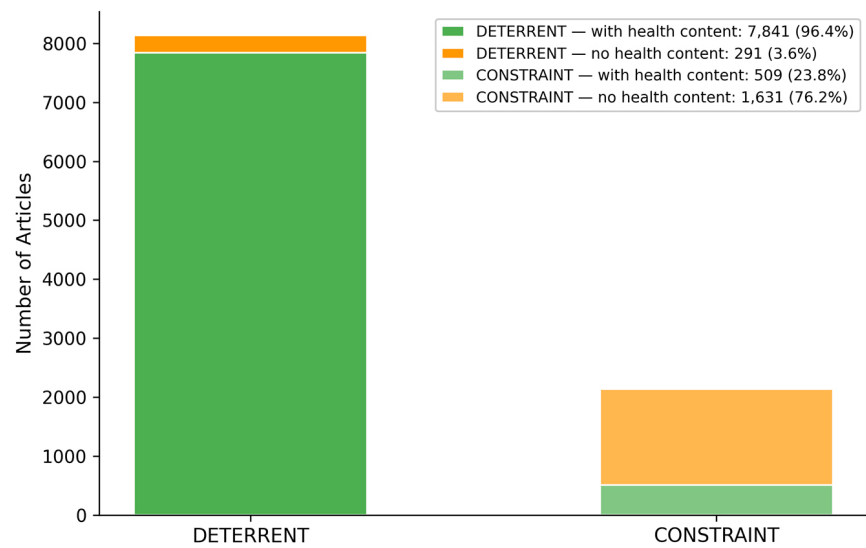


Fig. 4. Summarization results across both datasets

3.4 Claim extraction

The main goal of the claim extraction stage is to decompose each summary into a set of claims containing enough context to be understood and verified without the original article's content or title. Therefore, the LLM is prompted to extract claims respecting three constraints: (1) each claim must be a self-contained fact; (2) each claim must be atomic, meaning it should contain only one verifiable fact; and (3) each claim must preserve the exact wording and strength, sharing the same logic with the summarization stage described in Section 3.3. Apart from those instructions, to improve the reliability of LLM output further, few-shot examples are provided in the prompts.

Table 3 presents the claim extraction statistics, and Figure 5 shows the distribution of claims per article.

Table 3. Claim extraction statistics

Dataset	Articles	With Claims	Without Claims	Total Claims	Avg. Claims/Article
DETERRENT	8,132	7,829 (96.3%)	303 (3.7%)	72,601	9.3
CONSTRAINT	2,140	443 (20.7%)	1,697 (79.3%)	1,216	2.7

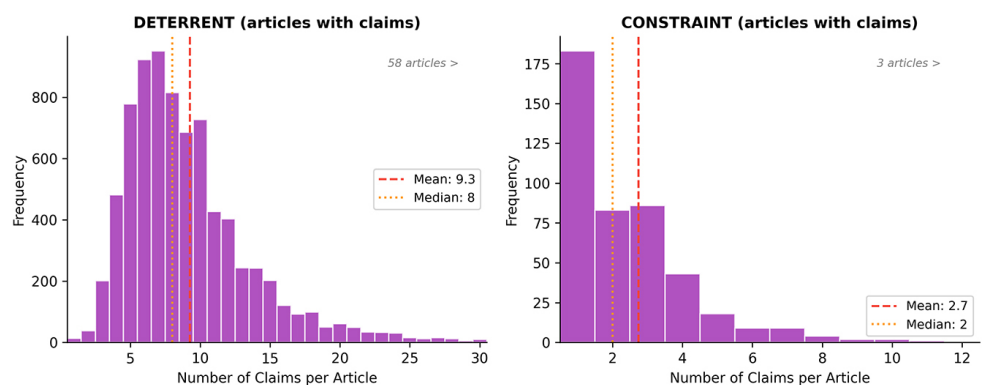


Fig. 5. Distribution of claims per article for DETERRENT and CONSTRAINT (only articles with at least one claim)

3.5 Knowledge base construction

Claims extracted from all entries in both datasets are vectorized using BGE-Large-en-v1.5, a 1024-dimensional dense embedding model. This embedding model was chosen for its strong performance on the Massive Text Embedding Benchmark (MTEB) [21]. The model produces a vector $\mathbf{v}_c \in \mathbb{R}^{1024}$ which is L2-normalized before being stored along with the reference label from the dataset:

$$\hat{\mathbf{v}}_c = \frac{\mathbf{v}_c}{\|\mathbf{v}_c\|_2}$$

where $\hat{\mathbf{v}}_c$ denotes the normalized embedding and $\|\mathbf{v}_c\|_2$ is the L2 norm of the original vector. Such normalization ensures efficient computation of cosine similarity between extracted claim c_i and evidence claim c_j from the knowledge base as the inner product of normalized embeddings:

$$\text{sim}(c_i, c_j) = \hat{\mathbf{v}}_{c_i}^\top \hat{\mathbf{v}}_{c_j}$$

Overall, for each claim, the following information is stored in the knowledge base: original claim text, original article's title and content, reference label (0 or 1), dataset source (DETERRENT or CONSTRAINT), and normalized embedding vector. By storing the dataset source, the system ensures that input claims are verified only against evidence from the same source. Another important design choice is storing both real- and misinformation-labelled claims, allowing the verification process (refer to Section 3.6) to compare each input claim against both categories of known evidence separately. Table 4 presents the final arrangement of claims in the resulting knowledge base.

Table 4. Knowledge base arrangement of claims by dataset and label

Dataset	Claims from Real Articles	Claims from Misinformation Articles	Total
DETERRENT	50,872	21,729	72,601
CONSTRAINT	736	480	1,216
Total	51,608	22,209	73,817

3.6 Claim verification

With the knowledge base in place, the verification process illustrated in Figure 3 can proceed. Initially, each input article passes through the same summarization and claim extraction stages as in the knowledge base construction process. Then, for each extracted claim from the input article, the system performs semantic retrieval of evidence and NLI verification.

Semantic retrieval. The extracted claim is embedded using the same BGE-Large-en-v1.5 model, making it possible to perform a similarity search. Given a query claim c_q and a label category $l \in \{0,1\}$, the retrieval returns the top- K evidence

claims, ordered by descending cosine similarity, whose value exceeds a minimum threshold τ .

$$R_l(c_q) = \text{top-}K\{c_r \in \text{KB}_l \mid \text{sim}(c_q, c_r) \geq \tau\}$$

where KB_l denotes the subset of knowledge base claims with label l , $K = 3$ is the number of retrieved claims, and $\tau = 0.83$ is the minimum similarity threshold. The threshold $\tau = 0.83$ is not intended as a strict matching criterion, since retrieved claims are still ranked and evaluated by the LLM regardless, but rather to filter out objectively unrelated candidates that could mislead an 8B-parameter model. Evidence retrieval is performed separately for each label category, each yielding three claims with the real label $R_0(c_q)$ and three claims with the misinformation label $R_1(c_q)$. In the case where no evidence claim in the knowledge base exceeds the similarity threshold, the input claim receives an automatic verdict of “not_enough_information” without invoking the LLM. To prevent data leakage, evidence claims extracted from the article being processed as input are dropped during retrieval, so they are never considered as candidates. This approach also addresses the issue of the lack of data-set splitting, as the LLM will not have any additional context.

LLM-based verification. Retrieved evidence along with the input claim is passed to the LLM, which is prompted to make a decision. The LLM should compare the input claim and evidence claims semantically, and if there is a “match,” it should determine whether the evidence supports or refutes the input claim. The prompt additionally instructs the LLM to evaluate the medical meaning rather than just structural similarity. For instance, if the input claim is about Vitamin C and the evidence claim is about Vitamin D, the LLM should yield a “no match,” even if the structure of both the input and evidence claims is similar. Once the LLM evaluates all retrieved evidence claims, it produces a final decision according to the following rules:

- The decision is “refuted” if any matched real evidence refutes the input claim, or if any matched misinformation evidence supports the input claim.
- The decision is “supported” if all matched real evidence supports the input claim and no misinformation evidence matches.
- The decision is “not_enough_information” if no evidence matches.

3.7 Plausibility assessment

Due to limitations in the self-constructed knowledge base, not all input claims have supporting or refuting evidence. To overcome this issue, an additional plausibility assessment is required of the LLM. The LLM is prompted to classify each claim based on its plausibility. “Consensus” claims are those that the model considers to be aligned with established medical knowledge, while “extreme” claims are those that the model considers to be opposed to established medical knowledge. Examples of “extreme” claims include miracle cures and unwarranted efficacy guarantees. Because this assessment is only a secondary signal, the LLM is instructed to flag a claim as “extreme” only when clearly warranted. Otherwise, the default classification “consensus” is assigned.

3.8 Verdict determination

In the final stage of the pipeline, claim-level results are aggregated into a prediction using a weighted scoring function. Let N denote the number of claims extracted from the input article; then the following ratios are computed:

- $r_{refuted} = N_{refuted}/N$, the fraction of input claims refuted by evidence;
- $r_{matched_false} = N_{matched_false}/N$, the fraction of input claims matching misinformation evidence;
- $r_{extreme} = N_{extreme}/N$, the fraction of input claims flagged as extreme and unsupported;
- $r_{supported} = N_{supported}/N$, the fraction of input claims supported by evidence.

These ratios are then used to calculate a misinformation score S_m and a real score S_r :

$$S_m = r_{refuted} \cdot w_{refuted} + r_{matched_false} \cdot w_{matched_false} + r_{extreme} \cdot w_{extreme}$$

$$S_r = r_{supported} \cdot w_{supported}$$

where $w_{refuted}$, $w_{matched_false}$, $w_{extreme}$ and $w_{supported}$ are weights reflecting the relative reliability of each evidence type. Their values were determined empirically and provided alongside their rationale in Table 5.

The final article-level prediction $\hat{y}$ is determined by the following decision function:

$$\hat{y} = \begin{cases} 1 \text{ (misinformation)} & \text{if } S_m > \theta_m \wedge S_m > S_r \\ 0 \text{ (real)} & \text{if } S_r > \theta_r \wedge S_r > \alpha \cdot S_m \\ \emptyset \text{ (abstain)} & \text{otherwise} \end{cases}$$

where θ_m and θ_r are minimum thresholds, and α is a dominance multiplier. The asymmetric multiplier ($\alpha = 1.2$) reflects the design principle discussed in Section 5, under which misclassifying misinformation as real is treated as the more critical error. Additionally, the misinformation threshold ($\theta_m = 0.4$) is set to capture such articles, so that even a modest concentration of refuted or matched-misinformation claims outweighs supporting evidence. The real threshold ($\theta_r = 0.2$) is intentionally lower than the others, because a small amount of support with no counter-evidence is enough to classify an article as real.

Table 5 summarizes the scoring parameters.

Table 5. Weighted scoring parameters

Parameter	Description	Value	Rationale
$w_{refuted}$	Weight for refuted claims ratio	2.0	Strongest misinformation signal: direct contradiction by KB
$w_{matched_false}$	Weight for matched misinformation claims ratio	1.0	Moderate signal: semantic overlap with known misinformation
$w_{extreme}$	Weight for extreme unsupported claims ratio	1.0	Moderate signal: plausibility-based, can be subjective
$w_{supported}$	Weight for supported claims ratio	2.0	Symmetric with refutation: direct confirmation by KB
θ_m	Minimum S_m to predict misinformation	0.4	Low threshold to maintain coverage
θ_r	Minimum S_r to predict real	0.2	Lower than misinformation: less evidence needed for real
α	Dominance multiplier for real prediction	1.2	Asymmetric: real must clearly outweigh misinformation

4 RESULTS

4.1 Evaluation metrics

To evaluate the presented system's performance five metrics are being used: accuracy, coverage, precision, recall, and macro F1. The **coverage** measures confident predictions only, meaning those articles for which the system is able to produce a label of either "real" or "misinformation". **Accuracy** is computed over the subset of articles that the system was able to classify. **Precision**, **recall**, and **macro F1** are also computed over covered subset of articles, where **macro F1** is the unweighted average of the two class-level F1 scores.

4.2 Per-dataset performance

The classification performance for each dataset and the verdict breakdown are presented in Tables 6 and 7, respectively.

Table 6. Per-dataset classification performance

Dataset	Articles	Confident	Coverage	Accuracy	Macro F1	Misinformation P	Misinformation R	Misinformation F1	Real P	Real R	Real F1
DETERRENT	8,132	7,697	94.7%	98.6%	0.980	95.1%	98.7%	0.969	99.6%	98.5%	0.991
CONSTRAINT	2,140	423	19.8%	94.8%	0.948	94.3%	96.0%	0.951	95.4%	93.5%	0.944

On DETERRENT, the system achieved 98.6% accuracy at 94.7% coverage, with a macro F1 of 0.980. This is consistent with the dataset's long articles yielding more verifiable claims. Misinformation recall reached 98.7%, with only 22 misinformation articles incorrectly predicted as real, representing 1.3% of the covered subset.

On the other hand, on the CONSTRAINT dataset, although the system accurately classified 94.8% of the covered subset, the overall coverage drops notably to 19.8%.

Table 7. Verdict breakdown per dataset

Dataset	Supported	Partial	Refuted	Matched Misinformation	Extreme	NEI (Abstain)
DETERRENT	2,552	3,372	551	1,220	2	435
CONSTRAINT	171	25	207	20	0	1,717

Another notable observation between the two datasets is revealed in the verdict distribution, as shown in Figure 6. Most DETERRENT articles are predicted real through partial or full support, reflecting the dataset's imbalance toward real articles. A substantial share of misinformation is detected through matched misinformation evidence and direct refutations. In CONSTRAINT, refutation dominates among confident predictions, likely because COVID-19-focused misinformation tends to contradict established medical knowledge directly.

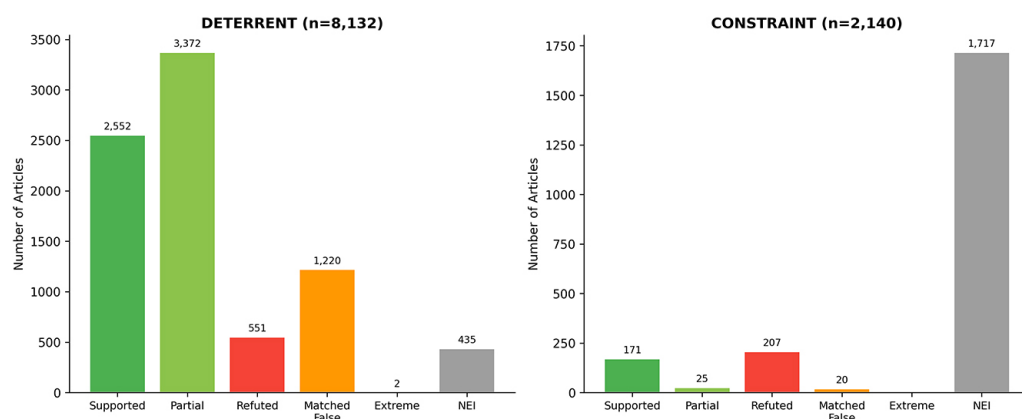


Fig. 6. Verdict distribution across datasets

The per-dataset confusion matrices over the covered subset of entries are presented in Figure 7, reflecting the asymmetric decision rules described in Section 3.8.

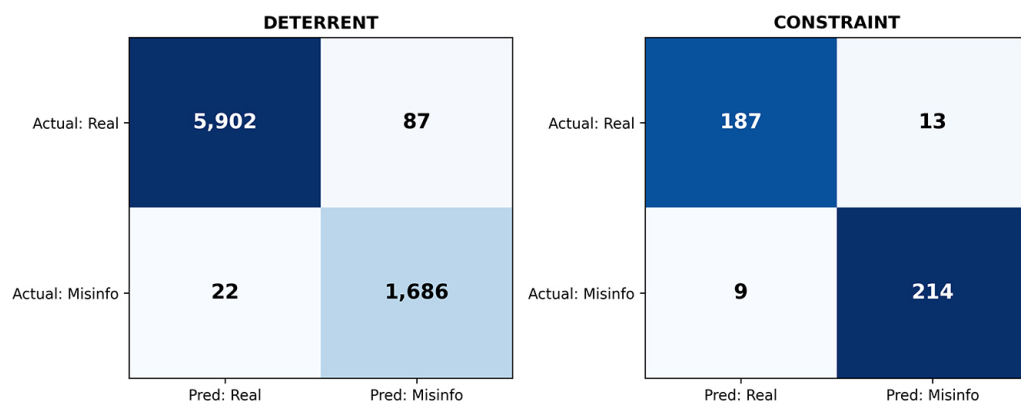


Fig. 7. Confusion matrices for both datasets

4.3 Aggregate performance

Beyond the per-dataset breakdown, Table 8 presents the aggregated report across both datasets.

Table 8. Aggregate classification performance

Articles	Confident	Coverage	Accuracy	Macro F1	Misinformation P	Misinformation R	Misinformation F1	Real P	Real R	Real F1
10,272	8,120	79.0%	98.4%	0.978	95.0%	98.4%	0.967	99.5%	98.4%	0.989

Overall, the system achieved a high accuracy of 98.4% and a macro F1 of 0.978 over the covered subset (79.0%). As shown in the aggregate confusion matrix in Figure 8, the system correctly identified 1,900 misinformation and 6,089 real articles, while only 100 misinformation and 31 real articles were misclassified. Therefore,

when the system decides that an article is real, it is almost always correct. Only a few misinformation articles can evade detection, provided they are in the covered subset.

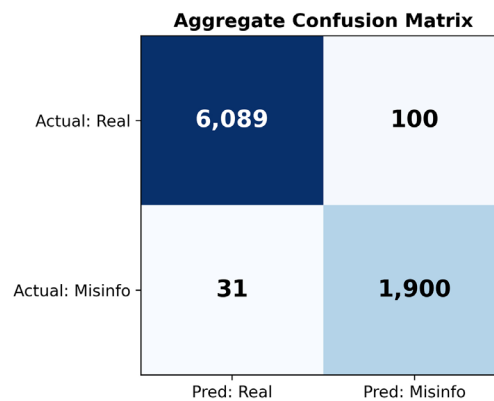


Fig. 8. Aggregate confusion matrix across both datasets

4.4 Coverage analysis

When the system is unable to extract at least one verifiable medical claim from an article, there is no basis for it to make a decision. This is a notable problem in the CONSTRAINT dataset. Short social media posts are less likely to contain any extractable claims, as shown in Figure 9.

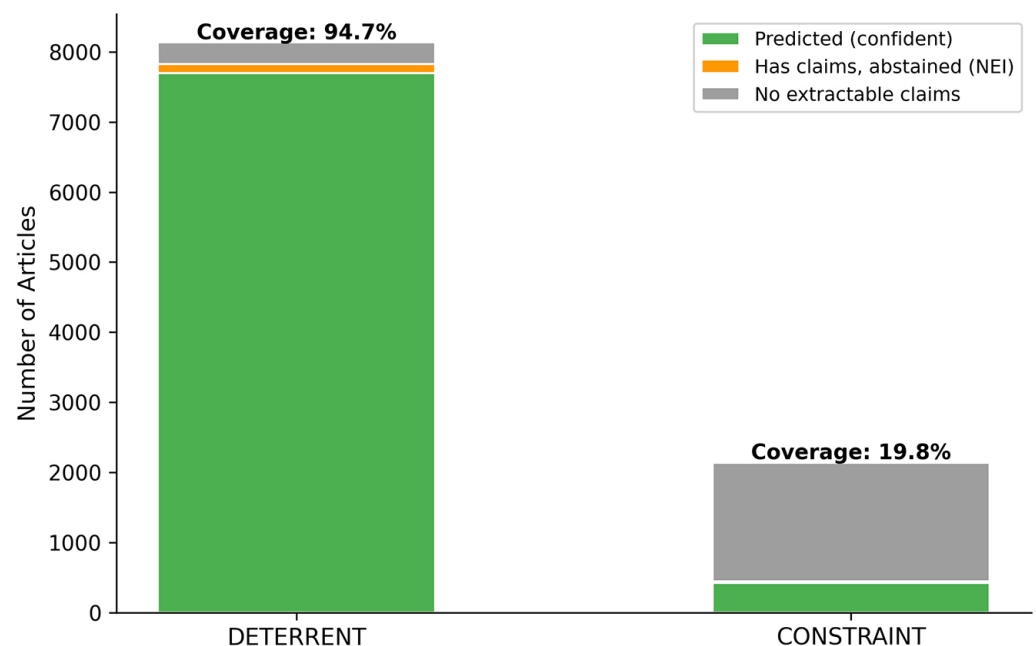


Fig. 9. Coverage breakdown per dataset

4.5 Ablation study

To further isolate the contribution of the proposed pipeline, an ablation experiment was conducted using the same LLaMA-3.1-8B model. The whole pipeline was replaced with a single prompt that classifies each article as either misinformation,

real, or unknown. This prompt was revised to achieve the best possible performance in this configuration. The same temperature parameter (0.1) and the same datasets were used. Table 9 compares the results of this experiment to the pipeline proposed in this paper, aggregated over both DETERRENT and CONSTRAINT datasets.

Table 9. Ablation: Single prompt vs. the proposed pipeline

Metric	Single Prompt	Proposed Pipeline	Δ
Coverage	90.0%	79.0%	−11.0pp
Accuracy	85.7%	98.4%	+12.7pp
Macro F1	0.773	0.978	+0.205
<i>Misinformation precision</i>	97.9%	95.0%	−2.9pp
Misinformation recall	46.9%	98.4%	+51.5pp
Misinformation F1	0.634	0.967	+0.333
Real recall	99.6%	98.4%	−1.2pp

The conducted experiment produced a relatively high coverage of 90.0%, but with a notably lower accuracy of 85.7%. Although this appears to be good performance, the misinformation recall reveals that the baseline configuration is unable to detect more than half of the misinformation entries. The high real recall (99.6%) is similarly explained by class imbalance. A trivial classifier that labels every article as “real” would already achieve 70.4% accuracy at 100% coverage.

5 DISCUSSION

The isolation of the contribution of the proposed architecture in this paper from high-performing LLMs and extensive knowledge bases is a key idea. The results presented in Section 4 demonstrate that even with a constrained setup, it is possible to achieve relatively competitive performance. This premise is supported by an additional ablation experiment, which shows that pipeline architecture is the main contribution. As a side note, no comparison with original evaluation datasets authors was made, since the proposed system differs. The published benchmarks on DETERRENT and CONSTRAINT report strong results using supervised classifiers trained on labels. For instance, a graph attention network achieves 0.847 to 0.931 F1 across DETERRENT’s diabetes and cancer subsets [6], and an SVM achieves 93.3% F1 on CONSTRAINT [7]. These results are not directly comparable to this paper’s, however. Both benchmarks are supervised and evaluate the full dataset, whereas the proposed system is unsupervised and only classifies articles for which it can retrieve sufficient evidence.

Given its nature, the system can be leveraged as a baseline for future research in the field. The fact that the system uses a self-constructed knowledge base further strengthens its role as a baseline. This makes it fully reproducible without needing additional access to external APIs.

Additionally, it is important to note another benefit of the self-constructed knowledge base. The knowledge base is extensible, since real-world content can be labelled by experts and added to it. Any system could integrate a pipeline that constantly

updates the knowledge base with new content. This would let it adapt to the changing nature of misinformation, likely improving coverage on datasets whose entries are long enough to yield verifiable medical claims.

Because the presented system checks the factual accuracy of specific claims, detecting misinformation in datasets such as CONSTRAINT would require additional processing. This could include semantic analysis or consideration of additional context. CONSTRAINT was nonetheless included intentionally, as an attempt to evaluate different formats of the entries. During the research, some design choices were explored to improve coverage on the CONSTRAINT dataset as well. One example is asking the LLM to semantically evaluate entries with no medical content. Although this showed some results during experiments, the final decision was not in favour of those changes. As stated previously, this is not the subject of the current study.

One of the important design principles of the system is to abstain from confident predictions if there is no factual basis. This same principle motivates the NO_HEALTH_CONTENT gate in article summarization (Section 3.3). Rather than forcing a classification on content it cannot verify, the system abstains from extracting claims entirely. It is considered worse to mislabel misinformation as real than to mislabel real content as misinformation. Both errors are worse than honestly stating there is no way to prove truthfulness. This asymmetry motivates the dominance multiplier used in verdict determination. It requires stronger evidence before predicting real than before predicting misinformation. Such abstentions should encourage users of the proposed system to look for additional, more reliable sources of information before deciding whether to trust the information. This is directly related to another key idea of this paper, which is to further improve the population's medical-demographic security.

6 LIMITATIONS AND FUTURE WORK

Despite these contributions, several limitations remain and suggest directions for future research.

Dataset diversity. The datasets used in this study already differ in format (long articles vs. short posts). Still, evaluating the approach on a broader and more multi-lingual range of datasets would further test its generalization. Most available misinformation detection datasets are also COVID-19-focused, so datasets on other topics would help test how well this or similar architectures generalize. The datasets used in this paper also limit the generalization of the self-constructed knowledge base.

Coverage. Semantic analysis of entries for which the system abstained is likely to further improve coverage, as discussed in Section 5. A study in this direction is already underway.

Model size. A relatively small language model was used to keep the approach applicable on limited hardware. The pipeline can already be reproduced without a graphical processing unit (GPU), though further optimizations are possible. One direction is to fine-tune smaller, raw language models, such as base T5, specifically for extracting verifiable medical claims.

Scoring parameters. The weighted scoring parameters were determined through empirical evaluation on the benchmark datasets used in this study, without validation on additional data. Replacing this stage with a traditional machine learning approach is one possible direction.

7 CONCLUSION

This paper proposed a modular retrieval-augmented verification system for detecting health misinformation. The pipeline combines knowledge base construction, article summarization and claim extraction, semantic retrieval of evidence claims, LLM-based NLI, and weighted verdict scoring. The system performed strongly on long-form content, achieving high accuracy and macro F1 on the DETERRENT dataset. Coverage on the CONSTRAINT dataset's short social media posts remained limited, however, reflecting the difficulty of extracting verifiable claims from brief text. These results show that a modular, claim-level verification pipeline can achieve competitive accuracy without depending on very large models or external knowledge sources. The proposed system is therefore both reproducible and accessible, offering a practical baseline for future work in health misinformation detection. By demonstrating that effective verification does not require resource-intensive infrastructure, this work contributes a concrete step toward protecting population health from the harms of medical misinformation.

8 ACKNOWLEDGMENTS

Gratitude is expressed to the Institute of Information Technology for its support of this study.

9 DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

The AI assistant Claude Sonnet 4.6 (Anthropic) was used for minor grammar polishing of the content. The final manuscript was written and revised by the authors, who take full responsibility for the content.

10 REFERENCES

- [1] V. Mammadaliyev and V. Shahbazov, "Disinformation detection in the medical domain: Current approaches, limitations, and future directions," *Problems of Information Society*, vol. 17, no. 1, pp. 81–94, 2026. <https://doi.org/10.25045/jpis.v17.i1.09>
- [2] World Health Organization, "Infodemics and misinformation negatively affect people's health behaviours, new WHO review finds," 2022. [Online]. Available: <https://www.who.int/europe/news/item/01-09-2022-infodemics-and-misinformation-negatively-affect-people-s-health-behaviours--new-who-review-finds>
- [3] M. M. Ferreira Cáceres *et al.*, "The impact of misinformation on the COVID-19 pandemic," *AIMS Public Health*, vol. 9, no. 2, pp. 262–277, 2022. <https://doi.org/10.3934/publichealth.2022018>
- [4] S. Loomba, A. de Figueiredo, S. J. Piatek, K. de Graaf, and H. J. Larson, "Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA," *Nature Human Behaviour*, vol. 5, no. 3, pp. 337–348, 2021. <https://doi.org/10.1038/s41562-021-01056-1>
- [5] H. I. Okagbue, P. E. Oguntunde, S. A. Bishop, E. C. M. Obasi, A. A. Opanuga, and O. P. Ogundile, "Review on the reliability of medical contents on YouTube," *International Journal of Online and Biomedical Engineering (ijOE)*, vol. 16, no. 1, pp. 83–99, 2020. <https://doi.org/10.3991/ijoe.v16i01.11558>

- [6] L. Cui, H. Seo, M. Tabar, F. Ma, S. Wang, and D. Lee, "DETERRENT: Knowledge guided graph attention network for detecting healthcare misinformation," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2020, pp. 492–502. <https://doi.org/10.1145/3394486.3403092>
- [7] P. Patwa *et al.*, "Fighting an infodemic: COVID-19 fake news dataset," in *Combating Online Hostile Posts in Regional Languages During Emergency Situation*, T. Chakraborty, K. Shu, H. R. Bernard, H. Liu, and M. S. Akhtar, Eds., vol. 1402, Cham, Switzerland: Springer, 2021, pp. 21–29. https://doi.org/10.1007/978-3-030-73696-5_3
- [8] C. Chen and K. Shu, "Combating misinformation in the age of LLMs: Opportunities and challenges," *AI Magazine*, vol. 45, no. 3, pp. 354–368, 2024. <https://doi.org/10.1002/aaai.12188>
- [9] E. Papageorgiou, C. Chronis, I. Varlamis, and Y. Himeur, "A survey on the use of large language models (LLMs) in fake news," *Future Internet*, vol. 16, no. 8, Art. no. 298, 2024. <https://doi.org/10.3390/fi16080298>
- [10] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, 2020, pp. 9459–9474.
- [11] H. Li, J. Huang, M. Ji, Y. Yang, and R. An, "Use of retrieval-augmented large language model for COVID-19 fact-checking: Development and usability study," *Journal of Medical Internet Research*, vol. 27, Art. no. e66098, 2025. <https://doi.org/10.2196/66098>
- [12] D. Quelle and A. Bovet, "The perils and promises of fact-checking with large language models," *Frontiers in Artificial Intelligence*, vol. 7, Art. no. 1341697, 2024. <https://doi.org/10.3389/frai.2024.1341697>
- [13] S. Min *et al.*, "FactScore: Fine-grained atomic evaluation of factual precision in long form text generation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 12076–12100. <https://doi.org/10.18653/v1/2023.emnlp-main.741>
- [14] Y. Wang *et al.*, "When helpfulness backfires: LLMs and the risk of false medical information due to sycophantic behavior," *npj Digital Medicine*, vol. 8, Art. no. 73, 2025. <https://doi.org/10.1038/s41746-025-02008-z>
- [15] Y. Gao *et al.*, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2023. <https://doi.org/10.48550/arXiv.2312.10997>
- [16] C. N. Hang, P.-D. Yu, and C. W. Tan, "TrumGPT: Graph-based retrieval-augmented large language model for fact-checking," *IEEE Transactions on Artificial Intelligence*, vol. 6, no. 11, pp. 3148–3162, 2025. <https://doi.org/10.1109/TAI.2025.3567369>
- [17] Y. Bai and K. Fu, "A large language model-based fake news detection framework with RAG fact-checking," in *Proceedings of the IEEE International Conference on Big Data*, 2024, pp. 8617–8619. <https://doi.org/10.1109/BigData62323.2024.10826000>
- [18] R. Upadhyay and M. Viviani, "Enhancing health information retrieval with RAG by prioritizing topical relevance and factual accuracy," *Discover Computing*, vol. 28, no. 1, Art. no. 27, 2025. <https://doi.org/10.1007/s10791-025-09505-5>
- [19] T. Hossain, R. L. Logan, A. Ugarte, Y. Matsubara, S. D. Young, and S. Singh, "COVIDLies: Detecting COVID-19 misinformation on social media," in *Proceedings of the 1st Workshop on NLP for COVID-19 at EMNLP 2020*, 2020. <https://doi.org/10.18653/v1/2020.nlpCOVID19-2.11>
- [20] J. Ma, W. Gao, S. Joty, and K.-F. Wong, "Sentence-level evidence embedding for claim verification with hierarchical attention networks," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019, pp. 2561–2571. <https://doi.org/10.18653/v1/P19-1244>
- [21] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "MTEB: Massive text embedding benchmark," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, Croatia, 2023, pp. 2014–2037. <https://doi.org/10.18653/v1/2023.eacl-main.148>

11 AUTHORS

Masuma Mammadova is a DSc, Professor and Corresponding Member of the Institute of Information Technology, Azerbaijan National Academy of Sciences in Informatics. She is Head of Department at the Institute of Information Technology, Ministry of Science and Education of the Republic of Azerbaijan, Baku, Azerbaijan. Her research interests include artificial intelligence, decision making, intelligent management systems, Health 4.0, digital medical twins, digital public health, and medical-demographic security (E-mail: mmg51@mail.ru).

Vagif Mammadaliyev is a PhD student at the Institute of Information Technology, Azerbaijan National Academy of Sciences, Baku, Azerbaijan. His research interests include medical-demographic security in the context of Medicine 4.0, and specifically the application of AI in the medical domain. He is the corresponding author of this paper (E-mail: v.mammadaliyev@iit.science.az).

Vusal Shahbazov is a PhD student at the Institute of Information Technology, Azerbaijan National Academy of Sciences, Baku, Azerbaijan. His research interests include information protection methods and systems, and information security in text-based content (E-mail: v.shahbazov@iit.science.az).