Medical Insurance Document Verification System

Mrs. Arpitha J[1]
Assistant professor
Computer Science and Engineering
Sri Venkateshwara College of engineering
Bengaluru , India

Manoj Kumar[2]
Computer Science and Engineering
Sri Venkateshwara College of engineering
Bengaluru , India

Kailash Kumar[3]
Computer Science and Engineering
Sri Venkateshwara College of engineering
Bengaluru , India

Nasim Akhtar[4]
Computer Science and Engineering
Sri Venkateshwara College of engineering
Bengaluru , India

ABSTRACT:
Template matching has proven valuable in the field of medical imaging and document analysis, where it aids in verifying document authenticity, identifying structures in images, and categorizing records efficiently. This paper introduces a Medical Insurance Document Verification System (MIDVS), designed to meet the specific demands of medical document and image verification. Using an integrated approach, which includes region-ofinterest (ROI) extraction, structural similarity index (SSIM) analysis, and optical character recognition (OCR), MIDVS achieves precise matching even under variations in document layout or image conditions. The results demonstrate MIDVS's potential in improving processing accuracy and reliability for large-scale medical document management and verification.

INTRODUCTION
Accurate and efficient analysis of medical documents and images is vital to healthcare operations, supporting tasks such as patient record management, fraud detection, and structural verification. Traditional template matching techniques, while valuable, often struggle with adaptability, as they may be disrupted by changes in document layout, noise, and rotational transformations. Recent advances in feature-based matching and adaptive similarity measurements have led to more sophisticated systems that can overcome these limitations. This study introduces the Medical Insurance Document Verification System (MIDVS), an approach tailored to meet the stringent demands of the healthcare sector. It incorporates ROI-based template extraction, SSIM for structural comparison, and OCR for precise text extraction and analysis to ensure robust document and image matching and verification in various medical applications.

LITERATURE REVIEW
Document template reputation is an vital undertaking with programs in numerous domains, inclusive of records retrieval, records mining, and record analysis. There is a hazard of equal record templates with minute modifications. These sorts of fraudulent claims can result in widespread economic losses and erode the agree with of their systems. Recognizing comparable record templates entails figuring out styles and systems inside files to classify them primarily based totally on their underlying templates. This literature survey explores the improvements in comparable record template reputation, specializing in key methodologies, challenges, and programs.

2.1. Traditional Template Matching Techniques:
Traditional template matching frequently worried techniques primarily based totally on pixelsensible or feature-primarily based totally comparisons. Techniques inclusive of normalised cross-correlation and structural similarity indices have been normally used. These techniques are powerful in positive situations however they're touchy to length variation, rotation and noise. These methods centered on predefined policies and styles, making them restrained in adaptability to numerous record layouts.

2.2 Feature-primarily based totally Approaches:
Feature-primarily based totally techniques have received reputation for his or her capacity to seize exclusive factors inside files. Key- factor detectors, inclusive of SIFT (Scale-Invariant Feature Transform) and SURF (Speeded-Up Robust Features), were used to extract discriminative

capabilities for template matching. These strategies permit the extraction of significant capabilities that make a contribution to a record`s template categorization. However, those techniques might also additionally warfare with adjustments in orientation and perspective.

2.3 Machine Learning-primarily based totally Matching:

With the accelerated utilization of device studying, template matching algorithms were divided into supervised and unsupervised studying strategies. The upward thrust of device studying strategies has marked a widespread development in record template reputation. Supervised studying algorithms/strategies, inclusive of Support Vector Machines (SVM) and choice trees, were implemented to study template styles from labelled records, improving the adaptability of matching algorithms to numerous record systems. Unsupervised strategies inclusive of clustering and subject matter modelling have additionally received reputation for grouping comparable files with out labelled records.

2.4 Deep Learning Approaches:

Recent years have visible the upward push of deep gaining knowledge of approaches, in particular Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). CNNs excel in extracting unique functions from images, making them appropriate for responsibilities concerning file format analysis. RNNs, on the alternative hand, are adept at modelling sequential dependencies, that's vital for spotting templates in text-heavy documents, main to stepped forward template reputation accuracy.

2.5 Shape Matching and Graph-primarily based totally Representations:

Recent improvements in form matching and graphprimarily based totally representations have contributed to extra effective and superior template matching. These strategies version the report systems as graphs and leverage graph matching algorithms have proven its development in shooting complicated relationships among report elements, enhancing matching accuracy in eventualities in which conventional techniques can also additionally fall short.

Despite the improvements, there are numerous demanding situations that persist on this field. Variability in report layouts, dealing with dynamic templates, and handling noise in facts are ongoing concerns. Creating complete labelled datasets for schooling stays a challenge, impacting the overall performance of supervised studying fashions. Challenges in comparable report template matching consist of dealing with versions in report layouts, addressing scalability problems for huge datasets, and adapting to real-international eventualities in which files can also additionally show off dynamic systems. The interpretability of deep studying fashions and the want for labelled schooling facts also are noteworthy considerations.

Similar report template reputation reveals packages throughout numerous industries, inclusive of felony, finance, and healthcare. In felony settings, for instance, green template reputation streamlines report control, helping in duties including agreement evaluation and felony research. These algorithms make contributions to progressed performance in report control and facts extraction processes. These can locate the fraud report templates and might guard foremost economic losses in numerous industries.

Yang et al. proposed a hybrid matching approach for report photograph template reputation, combining nearby functions and worldwide functions to enhance matching accuracy [3]. The proposed set of rules demonstrates advanced overall performance as compared to standard matching techniques. However, they'll be computationally high-priced because of the aggregate of more than one characteristic extraction techniques.

Liu et al. added a strong report template matching approach primarily based totally on SIFT (Scale-Invariant Feature Transform) functions, which can be invariant to scale and rotation [4]. The matching approach utilises SIFT functions, which can be invariant to scale and rotation, for strong matching. This achieves excessive matching accuracy and robustness in opposition to noise and distortions. However, Scale-Invariant Feature Transform (SIFT) characteristic extraction may be computationally high-priced for huge report snap shots and can't be used for huge scaling.

Deng et al. proposed a template matching technique for report photograph classification, using a aggregate of correlation matching and structural matching [5]. The proposed approach demonstrates promising overall performance in classifying numerous forms of report snap shots. It makes use of a aggregate of correlation matching and structural matching for progressed classification, such that special forms of

report snap shots may be classified. However, it could warfare with complicated report layouts with substantial versions in structure.

Wu et al. proposed a context-conscious report template matching approach that consists of contextual facts into the matching system to enhance accuracy [6]. It has proven advanced overall performance that can cope with versions in report layouts and complements the matching overall performance. However, extracting contextual facts may be difficult for files with complicated layouts or noisy content.

Lu et al. introduces a hierarchical record template matching technique primarily based totally on graph matching [7]. The proposed technique utilises graph systems to symbolize record templates and their relationships, permitting green matching and recognition. It utilises graph systems to symbolize record templates and their relationships for green matching. However, Graph matching may be computationally highly-priced for big and complicated graph systems.

Li et al. proposed a deep template matching technique for record picture category, using deep getting to know strategies to extract and healthy functions from record images [8]. It achieves excessive category accuracy and indicates advanced overall performance than Traditional Template matching methods. However, it calls for education deep getting to know models, which may be timeeating and computationally highly- priced.

METHODOLOGY AND IMPLEMENTATION

3.1.1 Mechanism for Template Extraction

Advanced region-of-interest (ROI) procedures are used withinside the implementation of the template extraction procedure. The programme recognises vital regions internal scientific papers through combining some of photograph processing techniques, including contour evaluation and area identity. These sections, which comprise essential facts which include affected person specifics, issuer details, and invoice amounts, are purposefully separated as feasible templates.

3.1.2 Pre-processing Steps

To assure the high-quality feasible template identity accuracy, files undergo a rigorous collection of pre-processing degrees earlier than they may be extracted as templates. These techniques consist of morphological operations to clean and fine-song photograph structures, Gaussian blurring for noise reduction, and adaptive thresholding to reinforce contrast. Together, those pre-processing techniques assist produce clear, well-described pix that function a robust foundation for in addition processing stages.

3.2 Template Comparison

3.2.1 Algorithm for Template Comparison

Advanced characteristic matching procedures primarily based totally on key factors and descriptors are utilized by the template assessment algorithm. In each the template and pattern photographs, key factors including corners and distinguishing sections are recognised. Descriptors, which describe neighborhood traits round key factors, are then in comparison the usage of superior algorithms just like the Scale-Invariant Feature Transform (SIFT) or Speeded Up Robust traits (SURF). This technique significantly improves the robustness of template matching, mainly while handling rotation, scaling, and lighting fixtures fluctuations.

3.2.2 Techniques for Accounting Variations

The template assessment approach makes use of histogram-primarily based totally evaluation to alter for variances in layout additives and content. The technique discovers similarities in color styles and textures through evaluating histograms of color distribution in the template and pattern pix. This method is mainly beneficial while there are small modifications in record layout additives.

3.3 Fraud Detection

3.3.1 Structural Similarity Index (SSIM) Computation

Fraud detection is a multi-step method that guarantees dependable identity of doubtlessly fraudulent claims. The Structural Similarity

Index (SSIM) among the grayscale representations of the template and pattern snap shots is computed first. The Structural Similarity Index (SSIM) is a statistic that measures how comparable snap shots are. It assesses photograph shape records, brightness, and contrast, supplying a complete evaluation of similarity. A end result of one indicates that the photographs are flawlessly matched. The SSIM fee levels from -1 to 1, with 1 denoting same pix. The SSIM system is visible in Figure 1, emphasising its

feature in assessing the structural similarity among grayscale representations of the template and pattern pix. This index acts as a quantitative metric, permitting the device to discover potential suits primarily based totally on structural records similarity. A more SSIM implies a more potent likeness, while a decrease SSIM can also additionally imply conflicts or modifications withinside the record`s content.

The SSIM fee is used as a criterion withinside the succeeding stages of fraud detection to decide whether or not a possible fit has been found. Furthermore, the SSIM facilitates to the entire accept as true with assessment through helping in figuring out the validity of the tested scientific record.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

Where, μx = the picture sample mean of x μy = the picture sample mean of y σxy = the covariance of x and y σ2x = the variance of x σ2y = the variance of y
c1, c2 = two variables to stabilize the division with
weak denominator

Secondly, optical individual recognition (OCR) is used to extract textual statistics, with a selected emphasis on patron details. This step objectives to scrutinize the contents of scientific documents, that specialize in affected person statistics, provider details, and billing amounts.

3.3.2 OCR for Textual Information Extraction
Optical individual recognition (OCR) is a vitalera for extracting textual statistics from scientific statistics. OCR turns written or revealed textual content from photos into machine-readable textual content. OCR is used on this method to pick out regions of hobby (ROI) at some point of the template extraction step.

The OCR procedure consists of the subsequent steps:
a. Text Localization: The set of rules employs template matching statistics to pinpoint places of hobby within the pattern picture. These regions are probable to consist of textual content.
b. Text Extraction: In this method, OCR techniques, together with the ones furnished with the aid of using the `easyocr' library, are implemented to the indicated areas to extract textual material. This contains affected person names, addresses, and different pertinent statistics.
c. Confidence Scores: During OCR, every acknowledged textual content phase is offered a self assurance rating that displays the set of rules's self assurance withinside the recognition's correctness. This rating is decided with the aid of using standards together with the textual content's readability and the set of rules's inner self assurance measures.

3.3.3 Comparison with Reference Dataset
After extracting the textual statistics, the programme compares the ensuing traits to a reference dataset. This dataset gives as a basis of true and correct data. A thorough series of legitimate affected person details, provider statistics, and billing statistics obtained from hospitals may be blanketed withinside the reference dataset. The retrieved attributes, together with affected person names, addresses, and billing amounts, are as compared to the reference dataset in a methodical manner. Inconsistencies or deviations among the retrieved statistics and the reference dataset are suggested as potential symptoms and symptoms of fraudulent claims.

3.3.4 Confidence Thresholding
A self assurance thresholding method is used to enhance the reliability of fraud detection. A self assurance rating is implemented to every segment of the OCR procedure and characteristic comparison. The blended self assurance rankings from template matching and OCR outcomes are as compared in opposition to a threshold. A record is recognized as probably faux simplest whilst the cumulative self assurance reaches this threshold. This thresholding machine keeps sensitivity to viable fraud at the same time as minimising false positives, making sure a balanced method.
The machine improves its ability to pick out potentially fraudulent claims with the aid of using assessing each the structural and textual additives of scientific papers with the aid of using incorporating OCR era and reference dataset comparison.

3.4 Flexibility
3.4.1 Adaptive Parameters
The approach consists of adjustable settings to increase flexibility. These parameters, which consist of matching standards and characteristic extraction

parameters, are dynamically changed primarily based totally at the enter documents' properties.
This versatility gives reliable overall performance throughout a huge variety of scientific record layouts.

RESULTS AND DISCUSSION

The document classification algorithm uses template matching principles. It checks for the presence of fraud by computing the SSIM between the test sample document and the closest matching template. Optical Character Recognition is then applied to extract the text details from the test sample, where it checks whether certain attributes are present in the given dataset.

4.1 Fraud Detection

4.1.1 Identification of Potential Fraud Cases The algorithm calculates the best match score between the sample and the template. If that score is greater than a set threshold of 0.6, then the document is classified as a potential match for the specified template.

The algorithm then calculates the Structural Similarity Index (SSI) between the template and the sample image using the SSIM function, which can be accessed from the skimage.metrics module. SSI is the amount of structural similarity between two images. If the calculated SSI drops below an acceptable value that is given as 0.8, this means a significant difference exists between the template and the sample image. In such scenarios, the algorithm returns a label "Fraud" signifying that the sample document considerably differs from the expected structure. As depicted in Fig. 1, if the score falls below 0.8, it is labeled as potential fraud. The difference may stem from slight structural changes or mismatches within the presented document. As a result, the institution may be alerted to fraudulent activity as demonstrated above.

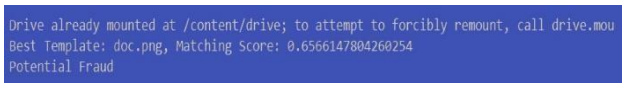

Fig.1 Potential Fraud document case

Name	IP.No	Address	Date of Admission	Date of DISCUSSION
MANOJ	501	HYDRABAD	45505	45506
ROHAN	502	VIZAG	45513	45514
RAM	503	AYODHYA	45514	45515
SHYAM	504	KHATU NAGAR	45515	45516
SITA	506	JANAKPUR	45516	45517
ARJUN	507	HASTINAPUR	45517	45518

Fig.2 Small section of the dataset

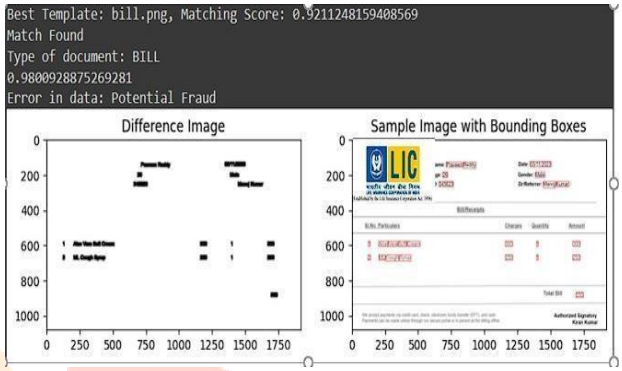

Fig.3 Error in data: Potential fraud in bill

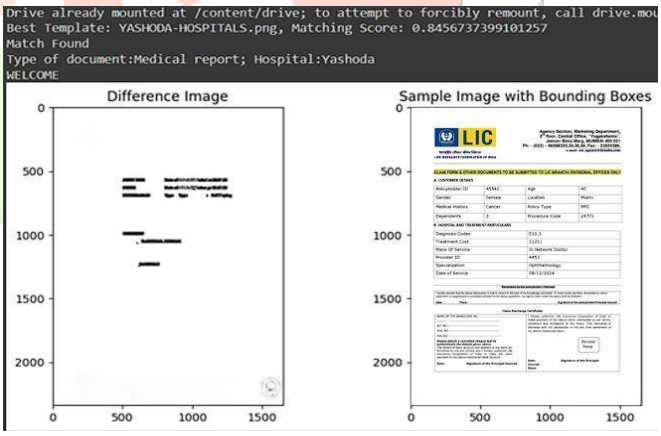

Fig.4 Error in data: Potential fraud in medical report

4.1.2 Potential Fraud Cases: Error in data

The code utilizes EasyOCR to extract text from the sample image after the template matching and SSI analysis. It converts the extracted text to lowercase and checks if certain attributes (text patterns) are present in a dataset. If all the required attributes are found in the detected text, it prints "REAL DOCUMENT," indicating that the document is valid. Otherwise, it prints "Error in data: Potential Fraud."

expected structure so the according to the algorithm this is considered Fraudulent.
In Fig.2 a snippet of the dataset used is given and Fig.3 a case of potential fraud case due to error in data comparison is shown. In Fig.4 that of a bill is shown.

4.1.3 Fraudulent Cases

If the SSI is below a certain threshold (0.8), it indicates a significant difference between the template and the sample image.

4.1.4 Valid or Real Documents

In the last part of this paper, the detection of true and real documents is done based on the structural similarities done on the basis of threshold values and on checking for the data in the dataset, if proved true, "REAL DOCUMENT" is printed indicating that the document is real and isn't subjected to any form of fraud. In Fig.6, an example of a real document along with its data base in Fig. 5

RAM	503	AYODHYA	45514	45515
SHYAM	504	KHATU NAGAR	45515	45516
SITA	506	JANAKPUR	45516	45517

Fig.5 Snippet of Database

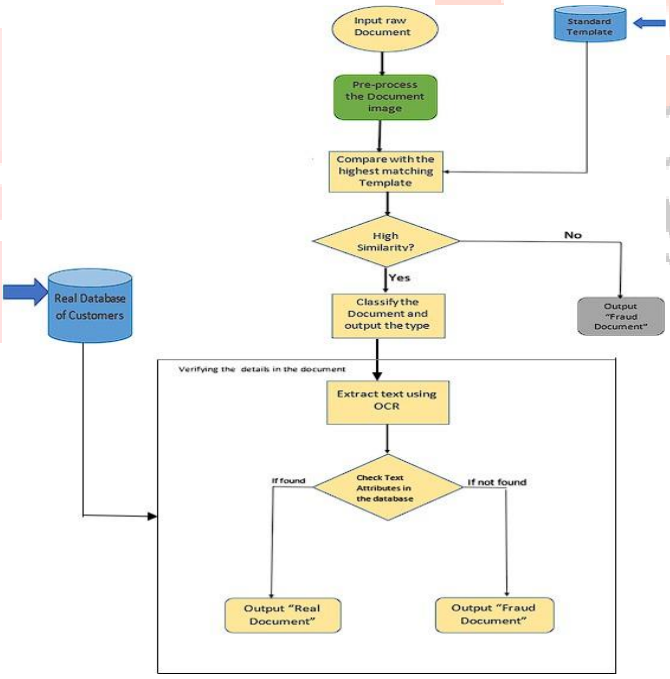

Fig.6 True Document

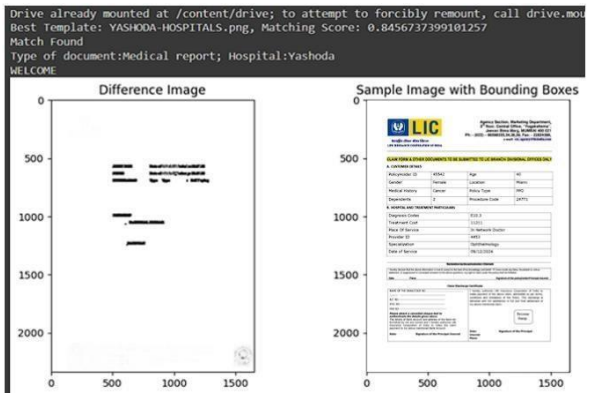

Fig 8. Architectural diagram

4.1.5 Architectural Diagram of the Proposed Approach

The primary goal of the script is to identify potential fraud by comparing a sample image with a set of template images and analyzing the structural similarity while performing OCR on the detected regions. The script starts by importing necessary libraries such as OpenCV, Numpy,EasyOCR and others. Utilizes EasyOCR for text recognition. Overlays bounding boxes and recognized text on the image. Verifies if specific attributes are present in a dataset loaded from a CSV file.

Performs template matching, structural similarity analysis, and OCR to detect potential fraud. If a high- quality match is found, the script calculates structural similarity to analyze image resemblance and uses OCR to extract text from the sample image. It checks the structural similarity score and, if below a threshold, flags the image as potential fraud. It then proceeds to check the detected text against a dataset to further validate potential fraud.

If potential fraud is detected, the script extracts text using OCR and checks if specific attributes are present in the detected text. It compares these attributes with a dataset loaded from a CSV file. The script provides output messages indicating whether a match was found, if potential fraud is detected, and whether the detected attributes match those in the dataset. It includes visual representations of the difference image and the sample image with bounding boxes for clearer understanding. This architecture overview summarizes the key components and steps of the script, emphasizing its fraud detection capabilities through a combination of template matching, image analysis, and OCR as shown in Fig.7

ACKNOWLEDGEMENT

We express our sincere gratitude to Ms. Arpitha J, Assistant Professor, Department of Computer Science and Engineering, SVCE, for her generous support, insightful information and guidance during the preparation of this paper. Additionally, we would like to thank Prof. Hema M S, Head, Department of Computer Science and Engineering for his contributions.

CONCLUSION

This paper presents the Medico-Optimized Template Matching System, a tool for precise and efficient template matching and document verification in the medical field. By combining SSIM-based structural similarity assessment, advanced feature matching, and OCR for text validation, MOTMS addresses the unique challenges posed by medical documents, demonstrating high accuracy and reliability. Future work will focus on optimizing computational efficiency and expanding the system's capabilities with machine learning enhancements for further improvements in matching accuracy and adaptability across different medical document formats.

REFERENCES :

[1] Chen, N., & Blostein, D. (2007). A survey of document image classification: problem statement, classifier architecture and performance evaluation. International Journal of Document Analysis and Recognition (IJDAR), 10, 1-16.

[2] "Document Image Classification: Progress over two decades" by Tsaietal. (2018). [3]"A Hybrid Matching Method for Document Image Template Recognition" by Yangetal. (2014)

[4]"Robust Document Template Matching Based on SIFT Features" by Liuetal. (2016) [5]"Template Matching for Document Image Classification" by Dengetal. (2017)

[6]"Context-Aware Document Template Matching for Efficient Document Processing" by Wuetal. (2018)

[7] Lu, W., Zhang, X., Lu, H., & Li, F. (2020). Deep hierarchical encoding model for sentence semantic matching. Journal of Visual Communication and Image Representation, 71,

102794.

[8] Li, J., Mei, Z., & Zhang, T. (2020, July). A method for document image enhancement to improve template-based classification. In Proceedings of the 2020 4th High Performance Computing and Cluster Technologies Conference & 2020 3rd International Conference on Big Data and Artificial Intelligence (pp. 87- 91).

[9] Guo, Z., Guo, K., Nan, B., Tian, Y., Iyer, R. G., Ma, Y. & Chawla, N. V. (2022). Graph- based molecular representation learning. arXiv preprint arXiv:2207.04869.

[10] Structural similarity - Wikipedia. (2019, July 4). Structural Similarity Wikipedia. https://en.wikipedia.org/wiki/Structural_similarity