Blockchain validation of condensed digitized PHR

***Abstract*— Secured storage of voluminous clinical records has posed challenges, especially with digital records gaining pace along with privacy concerns. Besides these the handwritten clinical records require digitization without losing crucial data. Such challenges were overcome using the following key technologies- NLP, OCR and Blockchain that were easily available, easy to setup and economically feasible. The collaboration of these domains has paved another firm step towards achieving the solution thereby helping patients, medical professionals in their day-to-day activities and better transparency. The Summarizer API responds to the POST requests with the summaries, when the middleware sends a request with the un-summarized clinical record. Similarly, the OCR API swiftly responds back with the digitized text from the handwritten records followed by the utilization of blockchain to validate, secure, and store the digitized, summarized clinical records for quick access during the Golden Hour, or any relevant situations.**

Keywords*— *REST API, Summarizers, Blockchain, Google Vision API, OCR, EHR, Clinical records.

I. INTRODUCTION

The recent surge in volume of medical records has spiked the need of better processing capabilities which was depicted in a report from Practo [1]. To handle such voluminous records, various technologies such as big data, blockchain, summarizers, translators, OCR (Optical Character Recognition), can be utilized. The project currently focusses on the combined power of blockchain and machine learning models to address the issues of privacy, security, transparency, digitization of PHR (Personal Health Records) and summarization of EHR (Electronic Health Record). The challenges in medical domain are not limited to the above problems, but also include quick retrieval of correct records in fraction of time especially during the Golden Hour where the patient needs the utmost quick medical care. Parsing through voluminous clinical records then, becomes very challenging irrespective of their forms- digitized or handwritten. Accompanied with this retaining the crucial details from the EHR and store them. This is achieved using the NLP (Natural Language Processing) Summarizers.

NLP is the art of interpreting human languages using mathematical models. Rahul et al. [2] have portrayed that there exist various sub domains comprising of various mathematical, statistical models in use to achieve purposes such as summarization, semantic analysis, generate textual content using document based, queries and generative AI (Artificial Intelligence). Such models can be utilized in various scenarios based on the creativity of the engineer to obtain magnificent outcomes. Summarizers can be classified based on various criteria, but the most chosen method of classification is based on the method of summarization that segregates them into two major classes – Abstractive and Extractive summarizers. The classification is based on the type of output they generate – in case of Abstractive they have the potential of generating novel words in the summaries whereas the Extractive summarizers choose to pick sentences from the given context and present them as their outputs without introducing any novelty in them. Models such as BERT (Bidirectional Encoder Representations from Transformers), BART (Bidirectional and Auto-Regressive Transformer) are classified as Abstractive Summarizers and models based on TF-IDF (Term Frequency - Inverse Document Frequency) concept such as Luhn, or other statistical models such as page rank algorithm are classified as Extractive summarizers as per these criteria. Technologies that closely revolves around the medicine domain is expected to be precise, quick, maintain semantic consistency, and be reliable. Based on the preliminary checks performed, it was concluded that the Extractive summarizers can maintain better semantic consistency as they rank the top 'x' number of sentences from the input to achieve the summary. The value of 'x' must be passed as a parameter in the function call of the respective functions. The data type must be of integer value only. In this process such models do not impose any new words nor remove words from sentences in turn resulting in summaries that retain some of the sentences as it is and hence the same semantics.

To brief about the dataset being used in this study comprised of Clinical records obtained from the website named- "MTSAMPLES.COM". The site claimed to consist of transcription of medical clinical records written by professionals during their interactions with the patients [3]. Various datasets were obtained by researchers for their study in this domain, such as – ICU (Intensive Care Unit) ward medical records were used by Emily Alsentzer et al. [4].

Looking forward to the digitization process, various ML (Machine Learning) models exists that claim to attempt digitization of hardcopies of text. Learning that there exist certain constraints such as time, money, and precision of digitization, it is equally important to acknowledge, incorporate sufficient details from the handwritten records. The Google's Vision API (Application Programming Interface) has been well known for maximum digitization with minimum noise in the output, has been implemented and discussed in the sections ahead.

The requirement of a well-known, accepted technology that securely stores, validates the records, and consistently performs the transactions is blockchain. G. A. Pierro et al. [5] introduces that there are various environments, frameworks that follow the concepts of majority voting and nodes. Ethereum, Solana are well known environments that enables the technology. Simpson et al. [6] has noted that with growth and evolution of effective management strategies, traditional methods have compromised with present day efficient security and data integrity. With similar studies, this research addresses such challenges with evolving blockchain technologies. This study also delivers a system which satisfies secured validation, decentralized storage, and transparent access control.

Leveraging the power of these key technologies, the problems have been addresses with key constrains as – economic feasibility, privacy, security, and consistency of clinical records. The subsections ahead describe in detail the procedure opted that enabled this opportunity to integrate blockchain and ML domains with the medical field.

II. LITERATURE REVIEW

In the recent past there have been many innovations, and research works. The relevant research works to this project have been complied below for the better understanding and extended learning.

Simpson et al. [6] presented a model for modern healthcare data management system. This module included patient and doctors' data validation and allowed patient centric approach which can brought a revolution in healthcare domain.

In a particular study by Bharimalla et al. [7], PHR were attempted to be converted to EHR and integrated blockchain into it. Many ML models were used such as CNN-LSTM (Convolutional Neural Network-Long Short-Term Memory) and Tesseract-OCR for the summarization and conversion of handwritten data to electronic form. Also, privacy was considered for the EHR and it was noted that the OCR mis predicted the characters at times.

R. Raj et al. [8] worked on blockchain applications in the pharmaceutical supply chain industry to improve transparency and security and combated counterfeiting.

A. Saini et al. [9] leveraged the blockchain and cloud storage technologies which reduced the network overhead through the judicious use of cloud storage. Security in cloud storage and scalability were presented as major concerns.

This paper by Chaves et al. [10] introduced a new way to summarize medical texts. It used two advanced networks, Convolutional Neural Networks (CNNs) and Conditional Generative Adversarial Networks (CGANs). Instead of the usual methods, this used an unusual way of picking sentences. It also includes a network derived from biology for better summarizing. Big changes included a new type of loss function that performed calculations. This change alone has increased performance by 5% on average, beating other models in tests on medical data. This method aims to make sure the grammar is correct and the meaning makes sense when summarizing from a single document in the CGAN structure.

The study by Rohil et al. [11] handled the swift increase in health-related writings. It suggested a summarization method that pulls together various elements. It used a mix of artificial intelligence and language processing to gather key data from patient health reports. Firstly, they created summaries from article abstracts. Then they used more complex process that combined medical term concept vectors with the Unified Modeling Language System. To form a packed summary, they utilized principles from graph theory. This method beats others when tested on PubMed MEDLINE articles. The aim is to effectively pull out and understand important information from the growing vault of health-related writings. This aided in the creation of smart health systems.

With increasing ever-present concerns of security and privacy, Hossain et al. [12] proposed the "HDM-chain" framework. This work can be effectively used for ensuring protection of data and allowing patients to control and overview their health data.

III. OVERVIEW

The study was divided into majorly two domains consisting of the ML (Machine Learning) and blockchain modules. Necessary steps were taken to ensure that the components function independently and hence were designed to be loosely coupled. To ensure that the services such as – Summarization and OCR run seamlessly, be maintainable and scalable they were designed as API (Application Programming Interface). In event of any disturbances or service upgrades they do not impact the overall system, storage of clinical records and the validation of records in blockchain.

As mentioned in introduction section, the Extractive Summarizers were found to be more appropriate for work, post implementation of various types of summarizers, and extensive literature survey. Models such as BART, XLNet, BERT, LSA, Luhn, NER (Named Entity Recognition), T5, and TextRank were implemented in Python and tested individually to rank their ROUGE (Recall - Oriented Understudy for Gisting Evaluation) score. To validate them the same input was passed to each of the above models to analyse thoroughly the outputs that were returned. Finally, the Summarization proceeded with only Extractive summarizers, due their higher ROUGE metric score. Wherein it was discovered that various models existed and worked on varying concepts of statistics. As mentioned earlier the Extractive summarizers need a parameter to produce the number of sentences a s output the 'x' value was set to 60% of the un summarised clinical records. This value originated post the trial of a range of values. Also, to note, the value of this 'x' can be modified as per requirements. The models finally chosen to be incorporated into the API were- Luhn that works based on the TF-IDF concept, TextRank on graph-based concept, and LSA based on SVD (Singular Value Decomposition). The pseudocode of the API is given below:

```
Approute("/")
Display welcome message
Approute("/rouge"), method= POST
Accept reference and hypothesis
Perform the ROUGE test
Return the statistics
Approute("/modelname"), method= POST
Accept the text
Perform the Summarization
Return the summarized text.
```

Following the implementation of the REST (Representational State Transfer) API, the most obvious question aroused if the summarizers were able to keep the crucial contents and how good was it performing. To address these queries the ROUGE score metrics were utilized to interpret and justify their performance form statistical point of view. But not to ignore the human evaluation is the key metric, as this ROUGE metric used n-gram concept to match words and phrases present in summaries to the actual un summarized clinical records, at times the context could be lost but not be measured. Hence to overcome this issue side by side human evaluation was also performed to ensure the quality. But the problem did not stop here, the API was tested locally but to let the other components access it, the API had to be made public. In the mean process Ngrok's cloud tunnelling service was discovered that enabled a secure tunnel between the Ngrok servers to the locally running API. Following which the API

was accessible publicly and the other components could access it using GET and POST requests. Interacting with the API was made very easy with JSON (Java Script Object Notation) format, that enabled interoperability of data amongst various components of the system.

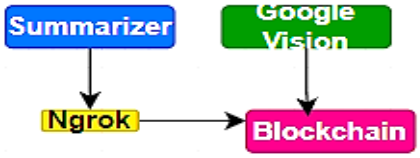

Figure 1.0: Depicting the interactions of various components of this study.

The Figure:1.0 depicts the arrangements and interactions of the components namely- Summarizer REST API, Google Vision API and Blockchain.

The OCR system has shown to be a useful tool for text extraction from images, supporting several image formats and carrying out the required preprocessing for improved image quality. Using Google's Vision API—which is excellent at converting text from images into a digital format; is now a necessary part of core OCR functionality. There exist many sophisticated features of Google's Vision API and the results are returned in JSON format. With strong error management and performance metrics, the system prioritizes data security when uploading images, and it improves user experience with a user-friendly guide and security features.

In the existing healthcare management systems, patients face several issues regarding their reports and diagnosis, such as unnecessary tests, irrelevant diagnoses, inaccessible medical files, and improper management of medical history and records. These issues result in inefficient diagnoses and excessive charges for medical procedures and treatment.

The developed solution aims to eradicate these issues and problems faced by patients. When patients visit the hospital for consultation, diagnosis, or any medical tests, the clinical records, or medical prescriptions (converted to digital texts using OCR Model) are uploaded to the web application through a simple interface by the hospital or doctors. These clinical records are then sent to a summarizer API. The response received from the API is stored on the blockchain, which can be viewed by patients after they log in to the application on their specialized dashboard. Additionally, the clinical text is sent to a registered medical professional who can view the patient's medical texts, add valuable insights, and validate the diagnosis done by the hospital and doctors. They can add comments against any issues or unrelated tests carried out by the hospital and doctors, making the patient aware and alerted about their tests and diagnosis.

To facilitate the above functionalities, the Ethereum testnet was chosen as the blockchain platform for its scalability, consensus mechanisms, and smart contract support. Prior to the use of the testnet, initial development and testing were carried out using the Ganache Local network. The major users of the application are doctors (or hospitals), patients, and medical professionals. To implement the essential functionalities for these users, four smart contracts were developed using Solidity, namely:

1. factoryContract (This smart contract contains methods for the registration of users and storing the references pointers to user contracts)

2. patientContract (This contains functions for retrieving patients' clinical texts, personal data, and providing access)

3. medicalProfContract (Contains methods for the addition of comments,and personal data)

4. doctorContract (Contains methods for addition patients and clinical text)

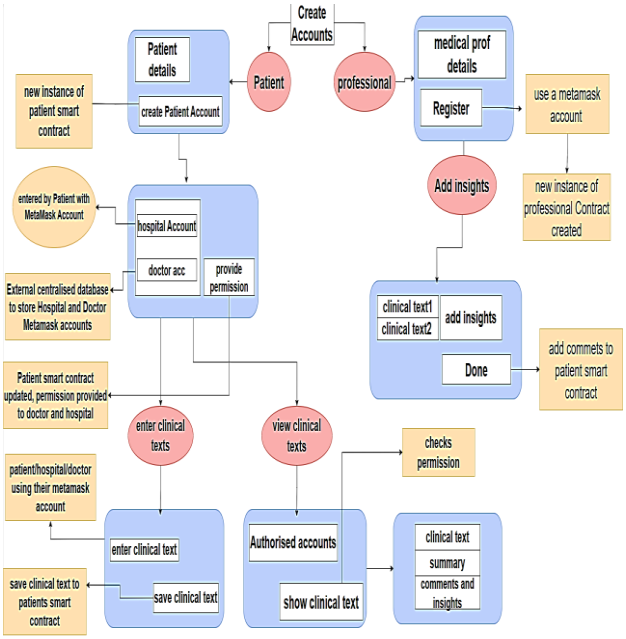

Figure 2.0: Depecting the the architecture to secure clinical summaries.

The Figure:2.0 demonstrates the complete architecture of the blockchain module of the system being used in this study. The interactions of the user, the transit of the data from one component to another and existence of features and functionalities working with components is shown in the above diagram.

At the beginning patients, doctors, and medical professionals must register on the application using their personal data through the UI (User Interface), which invokes functions dedicated to the registration of these users:

1. function createPatient(string memory _name, uint _age, string memory _birth, string memory _gender,string memory _phone){}. Creates a patient contract instance with the parameters saved into the contract.

2. function createDoctor(). Creates a doctor contract instance.

3. function createProf(string memory _name, uint _age, string memory _gender, string memory _hospital,string memory _hospitalId, string memory _email) {}. To create a medical professional instance with the parameters saved into the contract.

Users are required to be logged into their respective MetaMask accounts before registering themselves into the application. The same MetaMask accounts will be used in the future to log in to their registered accounts on the application. As the UI is designed with React.js and Next.js; Web3.js has been used to record and transfer the changes on the UI, to the blockchain and invoke the corresponding smart contracts.

After the registration, each user can log into their accounts using their MetaMask accounts. The application checks if the user is logged into their respective MetaMask accounts and takes the user to their personalized dashboard with just one click after successful authentication.

Both doctors and hospitals can add clinical records into the specialized interface using the unique patient id. When handwritten medical prescriptions were added, OCR API was called, and as a response, we received digitized text. Clinical records can be uploaded by doctors using the insertPatientDocs method. Which can be viewed by patients and medical professionals. Medical professionals can add comments and validate them using the addComments function in the medicalProfContract. These comments can be viewed by patients, making them aware of the authenticity of their diagnosis.

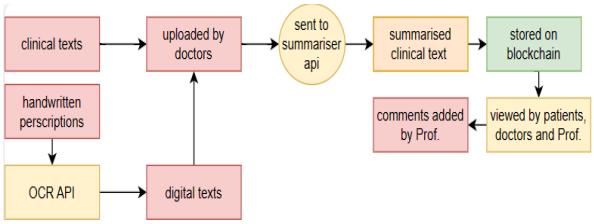

Figure 3.0: Demonstrating the sequence of steps taken to validate the clinical records.

The Figure 3.0 demonstrates the overall process or flow of the study to ensure the voluminous records are addressed accompanied by digitization of the PHRs, followed by the validation through the blockchain.

IV. Results and Discussion

The overall system was tested thoroughly and confirmed that the system does address the concerns raised in the introduction section and the detailed research of each component is presented in below.

Beginning with the Summarizer API, it was noted that the Luhn model achieved a ROUGE score of at least 68% whereas the TextRank models achieved at 71% ROUGE score, considering the 1-gram F1 scores only. It was also tested that the API could handle up to 1000 GET requests per second. It was also noted that the API responded back to the queries in less than 185ms. It was also observed that the response time was large when network congestion was high or the bandwidth was low for the WAN (Wide Area network) network. Hence concluding that the API was indeed performing well and was within the acceptable range of values as the minimum threshold was considered was 60% ROUGE score and the time of response per request was within reasonable range.

The following table demonstrates the ROUGE score for the same given input clinical record demonstrating the capabilities of the models used.

Table I: Summarizer models in REST API

Sno.	Model name	F1-score
1	Luhn	0.680
2	TextRank	0.743

Footnotes: Table demonstrates the models and their ROUGE scores

It was observed that the Google-'s Vision API is a better OCR compared to PyTesseract. It demonstrated great ability to recognize hand written text, hence it could be relied on. Appropriate security measures kept the user data safe when images were uploaded. A handy guide made it simpler to get text from photos. Such features have led to the decision of possible applications in the medical field. In contrast, Google Vision OCR performed better than PyTesseract OCR, with an amazing 97% accuracy rate compared to PyTesseract's 60%. Due to the priority of high precision in some applications such as this study, Google Vision appeared to be a reliable choice.

Moving on to the implementation of the blockchain, the component yielded promising results in terms of security, accessibility, and transparency. The use of Ethereum testnet as the platform provided a robust foundation, demonstrating scalability, efficient consensus, and comprehensive smart contract support.

The smart contracts designed for the system successfully facilitated secure interactions within the blockchain network. Functions dedicated to storing and retrieving patient data, implementing access controls, and validating medical records demonstrated reliable execution. The experimental feature allowed medical professionals to comment on patient's diagnoses and clinical records directly within the blockchain. This feature not only demonstrated functionality but also opened avenues for collaborative decision-making and knowledge sharing among healthcare professionals. The UI, designed with simplicity in mind, proved to be intuitive and user-friendly. The incorporation of Web3.js, React.js, and Next.js significantly influenced the effectiveness and user experience of the blockchain component within our healthcare data management system. The unique comments functionality showcased potential enhancements in patient care, with medical professionals contributing valuable insights to the diagnostic process. Following successful testing on the Ethereum testnet, the optimization of code for deployment on the Ethereum mainnet was addressed with scalability concerns and optimized gas usage. This transition ensured that the blockchain system is not only secure and transparent but also practical for real-world deployment.

To summarise, the results of the blockchain implementation showcase a secure and efficient healthcare data management system. The discussion highlights the potential impact on patient care through enhanced security, transparent data access, and collaborative decision-making among medical professionals.

V. Conclusion and Future Scope

The conclusion drawn from this study is that technologies can be utilized to aid in the medicine domain and in fact promotes more transparency, better reliability, and validation of the clinical records to prevent wrong medication.

Google's Vision API improved the effectiveness of automating the reading of patient reports and prescriptions from doctors. The goal of this study is to give healthcare professionals a dependable, useful, and effective tool so they can concentrate on giving excellent care instead of entering data. This technology promises for improving patient care, growth, and collaboration. The capabilities of this study have significantly improved with the integration of Google's Vision API.

To conclude, the integration of blockchain technology into the healthcare data management system has demonstrated significant advancements in data security, accessibility, and made the system more transparent for the patients. Looking forward, there are several aspects of the project which can lay foundation for the future research and improvements. Firstly, the system can be further optimized for deployment on the Ethereum mainnet, addressing potential challenges associated with real-world scalability and transaction costs. Secondly, introducing the feature such as explanation of difficult terminologies and appropriate medical explanation will make the system user friendly and will act as an essential helping hand to the patients.

The collaborative decision-making using comments on patient diagnoses can be refined, potentially incorporating machine learning algorithms to assist medical professionals in extracting valuable insights from accumulated data.

The Extractive summarizers have been in use for a long time for summarization, but with the advent of LLM (Large language Models) new opportunities have come where one can get the required work done using appropriate prompt engineering. Implying that the Extractive summarizers satisfied the requirements of the study, but with appropriate prompts this study can be extended further and explore the vast opportunities that lie ahead with summarization and generative AI (Artificial Intelligence) to achieve much better results.

VI. References

[1] ETHealthWorld. (2022, June 30). Nearly 60 per cent docs spent over eight hours per day on teleconsultations during pandemic: Practo Report.The Economic Times. https://health.economictimes.indiatimes.com/news/diagnostics/nearly-60-per-cent-docs-spent-over-eight-hours-per-day-on-teleconsultations-during-pandemic-practo-report/92570189.

[2] Rahul, S. Adhikari and Monika, "NLP based Machine Learning Approaches for Text Summarization," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2020, pp. 535-538, doi: 10.1109/ICCMC48092.2020.ICCMC-00099.

[3] https://www.mtsamples.com/index.asp

[4] Extractive Summarization of EHR Discharge Notes; Emily Alsentzer, Anne Kim; arXiv:1810.12085; https://doi.org/10.48550/arXiv.1810.12085.

[5] G. A. Pierro and A. Amoordon, "A Tool to check the Ownership of Solana's Smart Contracts," 2022 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER), Honolulu, HI, USA, 2022, pp. 1197-1202, doi: 10.1109/SANER53432.2022.00140.

[6] Simpson, G., Nana, L. and Kester, Q.A., 2021, December. A Centralized Data Validation System Model for Healthcare Systems Based on Blockchain. In 2021 International Conference on Cyber Security and Internet of Things (ICSIoT) (pp. 55-58). IEEE.

[7] Bharimalla, Pranab & Choudhury, Hammad & Parida, Shantipriya & Mallick, Debasish & Dash, Satya. (2022). A Blockchain and NLP Based Electronic Health Record System: Indian Subcontinent Context. Informatica. 45. 605-616. 10.31449/inf.v45i4.3503.

[8] R. Raj, N. Rai and S. Agarwal, "Anticounterfeiting in Pharmaceutical Supply Chain by establishing Proof of Ownership," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, India, 2019, pp. 1572-1577, doi: 10.1109/TENCON.2019.8929271.

[9] A. Saini, Q. Zhu, N. Singh, Y. Xiang, L. Gao and Y. Zhang, "A Smart-Contract-Based Access Control Framework for Cloud Smart Healthcare System," in IEEE Internet of Things Journal, vol. 8, no. 7, pp. 5914-5925, 1 April1, 2021, doi: 10.1109/JIOT.2020.3032997.

[10] Chaves, A.; Kesiku, C.; Garcia-Zapirain, B. Automatic Text Summarization of Biomedical Text Data: A Systematic Review. Information 2022, 13, 393. https://doi.org/10.3390/info13080393.

[11] Rohil, Mukesh Kumar & Magotra, Varun. (2022). An exploratory study of automatic text summarization in biomedical and healthcare domain. Healthcare Analytics. 2. 100058. 10.1016/j.health.2022.100058.

[12] Hossain, M.J., Wadud, M.A.H. and Alamin, M., 2021, November. Hdm-chain: A secure blockchain-based healthcare data management framework to ensure privacy and security in the health unit. In 2021 5th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT) (pp. 1-6). IEEE.