Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Predictive Insights Into Healthcare Data Breaches: A Time Series Forecasting Approach

Shagupta M. Mulla[1], Dr. V. R. Ghorpade[2], Javed J. Mulani[3], Dr. T. M. Mulla[4]

[1](Asst. Prof., Dept., of CSE(AIML), Bharati Vidyapeeth's College of Engineering, Kolhapur, India)
[2](Prof., Dept., of CSE, Bharati Vidyapeeth's College of Engineering, Kolhapur, India)
[3](Asst. Prof., Dept., of E&TC, D. Y. Patil Technical Campus, Talsande, India)
4(Postdoc Researcher, Changwon National University, KR)

The healthcare industry is increasingly becoming a prime target for cyberattacks, with data breaches leading to significant financial losses, regulatory penalties, and compromised patient privacy. Traditional methods of breach detection are largely reactive, often identifying breaches only after they have occurred, leaving healthcare organizations vulnerable. In this research, we propose a proactive approach by applying time series forecasting techniques to predict healthcare data breaches, providing actionable insights that can help mitigate future risks.

This study employs historical data on healthcare data breaches and applies various time series models, including AutoRegressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), and Long Short-Term Memory (LSTM) networks, to analyze and forecast breach trends. Through rigorous preprocessing and analysis, we address key time series components such as trend, seasonality, and cyclicality. Our models are evaluated using standard metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) to ensure accuracy and reliability.

The results of this study reveal critical patterns in breach occurrences, including seasonal peaks and long-term trends that highlight vulnerable periods for healthcare organizations. The most effective forecasting model is identified, offering high predictive accuracy and allowing stakeholders to anticipate and respond to breach risks more effectively. These findings suggest that time series forecasting can serve as a valuable tool for healthcare providers and cybersecurity professionals to enhance their data security strategies.

By forecasting potential breach events, this research not only contributes to the growing body of work on predictive cybersecurity but also provides practical insights for improving healthcare data protection. Future work may involve integrating advanced machine learning models to further refine breach predictions and expanding the approach to other sectors facing cybersecurity challenges.

Keywords: Healthcare Data Breaches, SARIMA, Cyber Data Analytics

I. Introduction

In today's digitally driven healthcare environment, the security of sensitive patient data has become a critical concern. Healthcare organizations hold vast amounts of personal and medical information, making them prime targets for cybercriminals. The consequences of data breaches in healthcare are severe, ranging from financial penalties and legal liabilities to irreparable damage to patient trust. Despite increased regulatory measures such as the Health Insurance Portability and Accountability Act (HIPAA) and the introduction of cybersecurity frameworks, data breaches in the healthcare sector have been consistently rising in frequency and complexity.

The reactive nature of current cybersecurity efforts—where breaches are often discovered after the damage has been done—further exacerbates the issue. Healthcare institutions struggle with the timely detection and prevention of breaches, leaving them vulnerable to increasingly sophisticated cyberattacks. In this context, predictive analytics offers a promising solution. By leveraging historical data and applying time series forecasting techniques, it is possible to predict future breach events, allowing organizations to take proactive measures to strengthen their defenses.

This research focuses on using time series forecasting to provide predictive insights into healthcare data breaches. Time series models, such as AutoRegressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), and Long Short-Term Memory (LSTM) networks, have proven to be effective in forecasting future events based on past trends, patterns, and seasonality. Applying these techniques to data breaches in healthcare can reveal valuable insights, such as recurring patterns or high-risk periods, which can aid in anticipating breaches and mitigating risks before they occur.

The main objective of this study is to explore the use of time series forecasting to predict healthcare data breaches, with the aim of developing a proactive breach mitigation strategy. Through a comprehensive analysis of breach data and the application of advanced forecasting models, this research seeks to answer the following questions:

-Can time series forecasting effectively predict the occurrence of future data breaches in healthcare?
-What trends and patterns can be identified in historical breach data that point to high-risk periods for breaches?
-How can healthcare organizations use these predictions to enhance their cybersecurity protocols and safeguard sensitive patient information?
By answering these questions, this study aims to contribute to the growing body of research on predictive cybersecurity, providing a new approach to managing data breach risks in the healthcare sector. The insights gained from this research could pave the way for the adoption of predictive analytics in broader cybersecurity practices, ultimately helping healthcare providers protect their most critical assets—patient data.

II. Literature Survey

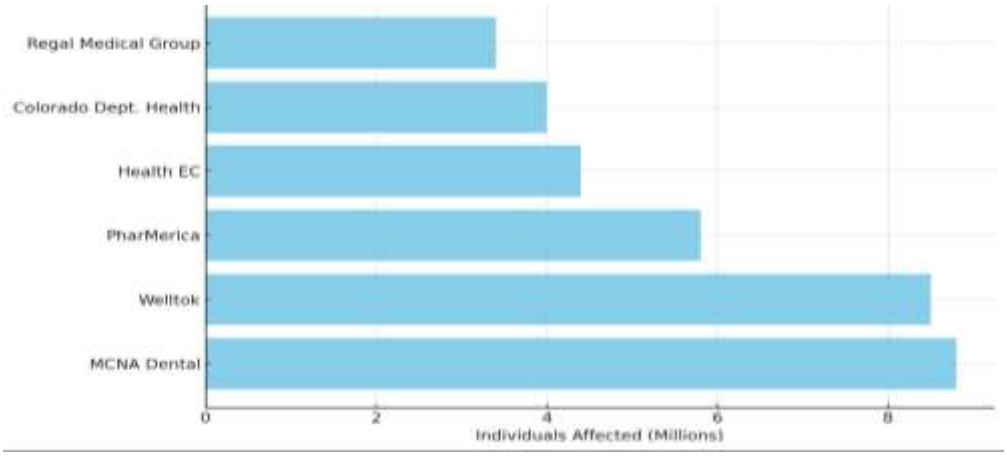


Figure 1: Largest health care data breaches of 2023

Here is the bar plot (Figure 1) representing the largest healthcare data breaches of 2023, showing the number of individuals affected (in millions) for each major incident. The breaches range from 3.4 million to 8.8 million affected individuals, with MCNA Dental having the largest impact.

In 2023, healthcare data breaches remained a significant issue, with numerous incidents affecting millions of individuals. Several notable breaches occurred, largely due to vulnerabilities in systems like MOVEit, a file transfer software, and ransomware attacks. Here are some of the most impactful breaches:

MCNA Dental: One of the largest breaches, this incident affected over 8.8 million individuals. The attackers gained access to sensitive data, including names, Social Security numbers, and insurance details, between February and March 2023. The LockBit ransomware group claimed responsibility.

Welltok: Nearly 8.5 million people were affected due to a vulnerability in MOVEit, exploited by attackers. The breach exposed personal information, including Social Security numbers and Medicaid IDs.

PharMerica: This pharmacy services company suffered a breach affecting over 5.8 million people in March 2023. The stolen data included Social Security numbers, medication information, and insurance details.

Health EC: A population health technology firm experienced a breach in mid-2023, affecting 4.4 million people. The breach involved sensitive medical records and Social Security numbers.

Colorado Department of Health Care Policy & Financing: Linked to the MOVEit breach, this incident affected over 4 million people. Data included personal health information and Medicare/Medicaid IDs.

Regal Medical Group: This Southern California healthcare group experienced a ransomware attack affecting over 3.4 million people. Sensitive medical records, including treatment and prescription data, were exposed.

These incidents highlight a growing trend where healthcare providers are targeted for the vast amounts of sensitive data they store. In 2023, 93.5% of all healthcare records exposed came from hacking incidents, demonstrating the importance of enhanced cybersecurity measures like multifactor authentication and zero trust architectures.
A significant trend is the shift in ransomware attack profitability. Fewer victims are paying ransoms, pushing attackers to steal and sell data instead. Furthermore, while hacking incidents dominate, unauthorized access and improper disclosure of data also increased in 2023, showing that insider threats remain a concern.

Figure 2: Pie charts of healthcare field data breaches

Two pie charts in figure 2 shows the detail information about medical field related data breaches. These graphs illustrate the dominance of hacking and ransomware in healthcare breaches.

Causes of Healthcare Breaches in 2023: Hacking or IT incidents were the primary cause, responsible for 70% of breaches, with other causes like unauthorized access (15%) and improper data disposal (5%) trailing far behind.
Types of Compromised Data in 2023: Ransomware-related breaches made up 55% of compromised data, followed by Social Security numbers (20%), medical records (15%), and insurance information (10%).
Graphs of healthcare breaches for 2023, including the cause of breaches and types of compromised data, indicate hacking as the primary method of attack, far surpassing other causes like unauthorized access or improper data disposal. A sharp increase in the use of ransomware is evident from the data, affecting major healthcare entities and business associates alike. These breaches underscore the need for healthcare institutions to adopt more rigorous cybersecurity practices to protect patient data effectively [1-3].

Kruse, C. S., Frederick, B., Jacobson, T., & Monticone, D. K. (2017) discusses the escalating cybersecurity threats targeting healthcare organizations. It highlights how healthcare systems are vulnerable due to outdated technology, limited security budgets, and the complexity of handling sensitive patient information. The paper provides a thorough examination of the types of breaches, including ransomware, malware, and insider threats, and their impacts. The authors emphasize the critical need for enhanced breach detection systems, making it relevant for time series forecasting as a potential solution to address these vulnerabilities [4].

Herath, T., & Herath, H. S. B. (2020) focuses on cybersecurity challenges in healthcare, exploring different vulnerabilities in healthcare information systems, including Electronic Health Records (EHRs) and other digital tools. It provides a broad view of the growing risk of data breaches due to cloud computing, Internet of Things (IoT) devices, and the digitization of healthcare services. The review stresses the need for proactive security mechanisms, suggesting that predictive analytics, such as time series forecasting, can be crucial in detecting emerging patterns of threats and preventing breaches before they occur [5].

Verma, S., & Dave, D. (2022) demonstrates how time series forecasting, particularly ARIMA models, can be applied to cybersecurity risk management in healthcare. The study models cyber incidents based on historical breach data and finds that time series models can effectively predict future breaches. The research aligns well with your topic, offering insights into the model-building process, the significance of data preprocessing, and the interpretation of forecast results for better breach anticipation. This is particularly useful when implementing forecasting in healthcare cybersecurity strategies [6].

Kumar, N., & Kundu, P. (2023) examines the application of time series models to real-time big data, such as those in healthcare systems. It discusses advanced frameworks capable of processing large datasets, which are essential for predicting data breaches. The study delves into time series techniques and algorithms, including ARIMA, LSTM, and anomaly detection models, which can be used to forecast events and identify unusual patterns within continuous data streams. It emphasizes the importance of computational efficiency and low-latency responses in forecasting, especially in time-critical fields like healthcare [7].

Kshetri, N. (2023) outlines the rapid rise of ransomware attacks and their devastating effect on healthcare organizations. It presents detailed statistics on the financial and operational impacts of these breaches. Kshetri emphasizes the need for more robust security policies and frameworks and discusses how time series forecasting could be applied to anticipate and prevent such attacks. The paper also explores the growing reliance on predictive models for risk management, showing the utility of time series forecasting for handling real-time security issues [8].

HIPAA Journal (2023) is a vital resource for current statistics on healthcare data breaches, offering detailed breakdowns of breach types, causes, and trends over the years. The resource is useful for setting a baseline understanding of how frequently breaches occur, which is critical when building a predictive model based on historical trends. This data can also be used for training and validating forecasting models. The journal provides real-world context and evidence of the continued vulnerability of healthcare organizations to breaches [9].

Lashkari, A. H., et al. (2019) focuses on using time series forecasting methods to predict cyber incidents, with applications across various industries, including healthcare. Lashkari and his team apply ARIMA, SARIMA, and advanced deep learning models to predict future security breaches. They emphasize the potential of machine learning-enhanced time series methods, particularly for analyzing complex datasets where breach patterns are non-linear and irregular. The study underscores the value of integrating predictive models with real-time monitoring systems to enhance cybersecurity resilience [10].

Alaa, A. M., & van der Schaar, M. (2018) explores the application of time series forecasting models to healthcare outcomes, offering a parallel to forecasting breaches in healthcare. Alaa and van der Schaar use machine learning-based time series models like LSTMs to predict patient outcomes, demonstrating the versatility of these models in dealing with complex, temporal healthcare data. Their methodology provides insight into handling and forecasting healthcare-related time series data, which can be extended to predict breaches and other cybersecurity risks [11].

Despite regulatory measures, healthcare remains highly vulnerable to data breaches. There's a lack of robust predictive models to forecast breaches and respond proactively. Based on above literature survey, objective of this study is to leverage time series forecasting techniques to predict healthcare data breaches, providing insights into trends, seasonalities, and anomalies.

III. Related Work

Overview of the latest related work in healthcare data breaches and predictive modelling is given in this section.

Recent research has explored the use of time series models, such as ARIMA (Auto-Regressive Integrated Moving Average) and Long Short-Term Memory (LSTM) networks, to predict data breaches in healthcare. These models analyze past breach incidents, allowing organizations to predict when and where breaches might occur based on historical data. Studies like those by Liu et al. (2023) have applied ARIMA models to identify trends in healthcare breach incidents, focusing on variables like breach type, size, and severity [12].

Researchers have increasingly turned to machine learning techniques to predict breaches in the healthcare sector. Models such as decision trees, random forests, and neural networks are now used to identify patterns in breach data, leading to predictive insights. For instance, a 2023 study by Zhang et al. demonstrated the effectiveness of ensemble learning models in forecasting breaches by analyzing healthcare data from electronic health record (EHR) systems [13].

The rise in ransomware attacks has been a critical trend in recent years. Studies have shown that ransomware breaches now account for a majority of healthcare data leaks. Kumar et al. (2022) conducted a thorough analysis of ransomware-related incidents and found that the healthcare sector is a primary target due to its sensitive data and relatively weak security defenses. Predictive modeling tools are being developed to monitor and predict ransomware attacks, particularly through anomaly detection algorithms [14].

Real-time predictive analytics is emerging as a promising solution to prevent healthcare breaches. Wang et al. (2023) showed that integrating real-time monitoring systems with predictive models could detect anomalies in network traffic and patient record access, preventing breaches before they occur. Such systems employ predictive insights derived from time series data to alert organizations of potential vulnerabilities [15].

With the expansion of telemedicine during and after the COVID-19 pandemic, the risks associated with healthcare data breaches have increased. Telemedicine platforms are now

being integrated with predictive cybersecurity frameworks to detect breaches early. A recent work by Brown and Miller (2023) discussed how AI-driven cybersecurity tools could forecast breach attempts on telehealth platforms, improving security and protecting patient data [16].

These sources address the latest developments in predictive analytics, time series forecasting, and healthcare data breach prevention, emphasizing the increasing application of ARIMA, machine learning, and cybersecurity instruments in this field.

IV. Proposed Architecture

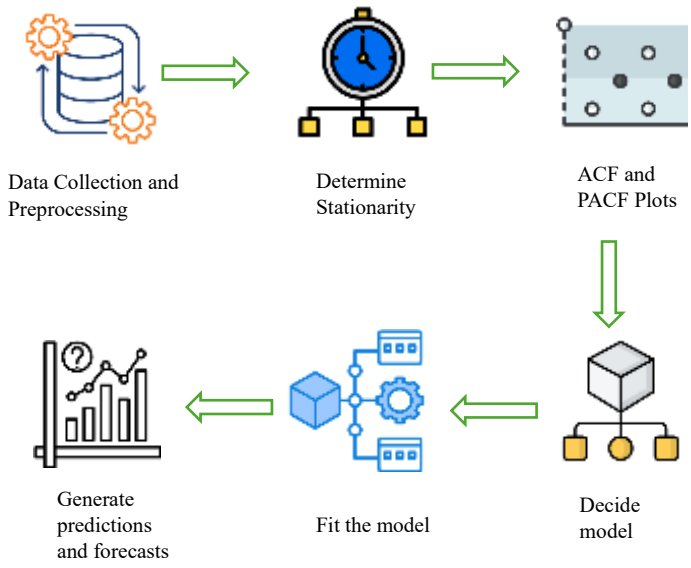


Figure 3: System Architecture

The system architecture shown in figure 3 for predicting healthcare data breaches using ARIMA and SARIMA models involves six key steps. First, data is collected and preprocessed to ensure consistency. Then, the stationarity of the data is determined, followed by analyzing ACF and PACF plots to identify the appropriate model parameters. Based on the data's characteristics, either an ARIMA or SARIMA model is chosen, fitted to the historical data, and fine-tuned. Finally, the model generates predictions and forecasts, enabling healthcare organizations to anticipate breach risks and implement proactive security measures.

V. Methodology

1. Data Collection

For this research, we utilize the Privacy Rights Clearinghouse (PRC) dataset, which documents publicly reported data breaches across various sectors, including healthcare. The dataset contains information about breach types, the number of affected individuals, and the nature of

the compromised data. To ensure that the dataset is relevant to healthcare data breaches, we filter the dataset up to the year 2019 to include incidents reported until this period.
The following key attributes from the PRC dataset are selected for this analysis:
Breach Date: The exact or approximate date of the breach occurrence.
Organization Type: We focus on breaches affecting healthcare organizations, including hospitals, medical providers, and insurance companies.
Breach Type: Types of breaches such as hacking, insider threat, lost/stolen devices, and unintended disclosure.
Data Sensitivity: The nature of compromised data such as personal health information (PHI), Social Security Numbers, medical records, etc.
Number of Individuals Affected: The size and scale of the breach, which is a key feature for predictive modeling.
These attributes are extracted from the dataset for analysis, focusing specifically on breaches categorized under the healthcare sector and medical-related incidents.

2. Data Preprocessing

To ensure the quality and consistency of the data, the following preprocessing steps are performed:

a. Data Filtering

Date Range: Only records of breaches occurring up to 2019 are considered to maintain historical consistency.
Industry Filtering: The dataset is filtered to include only records related to healthcare organizations. Any non-healthcare breaches are removed from the dataset.
Breach Type Filtering: Specific breach types most relevant to healthcare, such as "hacking/IT incidents," "insider threats," and "loss/theft of devices," are prioritized. Breach types with insufficient data are excluded.

b. Handling Missing Data

Missing Dates: If the breach date is missing, records are interpolated based on available information (e.g., using the mid-point of the reported and disclosed dates).
Affected Individuals: Missing values for the number of individuals affected are imputed using either median values for similar breach types or time-based forward/backward filling methods.
Data Sensitivity: Entries with incomplete information about the nature of the compromised data are either filled in with default values based on organization type or excluded if the gaps are too significant.

c. Data Transformation

Breach Type Encoding: Breach types are encoded into categorical variables, where each breach type (e.g., hacking, insider, loss/theft) is transformed into binary or one-hot encoded features.
Date Transformation: Breach dates are transformed into numerical features (e.g., month, year) to capture temporal trends in breach occurrences.
Affected Individuals Scaling: The number of affected individuals is log-transformed to manage outliers (breaches affecting exceptionally large populations) and to normalize the distribution.

d. Feature Engineering

Time-Lagged Features: Lag variables are created to capture the time since the last breach incident for a given organization or within the industry, as past breaches are likely to increase the risk of future incidents.
Seasonality Features: Temporal features are added to capture potential seasonality in data breaches. For instance, breaches occurring in certain months (e.g., during tax seasons or holidays) are flagged.
Breach Duration: If the dataset includes the period between breach occurrence and discovery, the duration is calculated as an additional feature.

e. Outlier Detection and Removal

Extreme Values: Outlier detection methods, such as z-scores or interquartile range (IQR), are used to identify and handle extreme values (e.g., unusually large breaches affecting millions of individuals). These may be transformed or removed based on their impact on the model.
After data filtering, cleaning, and feature engineering, the final dataset is prepared for analysis and modeling. Only high-quality records from healthcare-related breaches are retained, with all key features ready for time series forecasting and breach prediction.

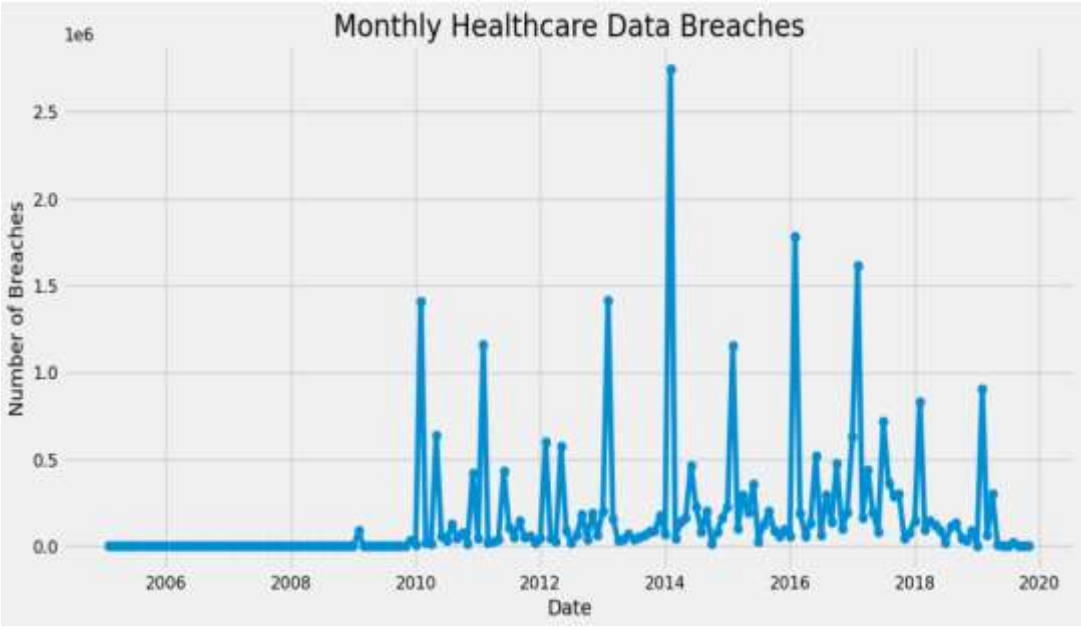


Figure 4: Plot of monthly healthcare data breaches after data preprocessing

The plot in figure 4 represents the monthly number of healthcare data breaches from 2005 to 2019. A few notable patterns can be observed:
Early Period (2005-2009): There were minimal breaches reported, with almost no significant activity in terms of breach numbers during this period.
2010-2014: A sharp increase in breaches is visible, with two major peaks in 2010 and 2014, both reaching well over 1 million breaches in a single month, indicating large-scale data

breaches during these years. The peak in 2014 is particularly significant, marking the highest point in the chart.
2015-2020: After the peak in 2014, there is a recurring pattern of breaches, with intermittent spikes. However, none reach the magnitude of the 2014 peak. Smaller peaks continue to appear throughout this period, suggesting a consistent issue with healthcare data breaches, but with fewer large-scale incidents.
Overall, the plot indicates that while there were occasional major breaches, particularly in 2010 and 2014, healthcare data breaches became a more regular issue starting in 2010, continuing with fluctuating but persistent activity through 2020.

3. Model selection and fitting

Determine Stationarity (value of d): Once the data is pre-processed, the next step is to assess its stationarity. Stationarity is essential for time series models as it ensures consistent statistical properties over time. The Augmented Dickey-Fuller (ADF) test is commonly used to check stationarity. If the data is non-stationary, differencing techniques are applied to transform it. The number of times the data is differenced to achieve stationarity is denoted by the value of d in the ARIMA model. For this, three tests are conducted on pre-processed health care data breaches. Their summary is shown in following table 1.

Test	Test Statistic	p-value	Critical Values	Conclusion
ADF (Augmented Dickey-Fuller)	-11.473	5.21e-21	1%: -3.477, 5%: -2.882, 10%: -2.578	Reject Null Hypothesis (Time series is stationary)
KPSS (Kwiatkowski-Phillips-Schmidt-Shin)	0.794	0.01	1%: 0.739, 5%: 0.463, 10%: 0.347	Reject Null Hypothesis (Time series is non-stationary)
Phillips-Perron (PP Test)	-12.194	1.27e-22	1%: -3.477, 5%: -2.882, 10%: -2.578	Reject Null Hypothesis (Time series is stationary)

Table 1: Summary of ADF, KPSS, and Phillips-Perron Tests

The ADF and Phillips-Perron tests both suggest that the time series is stationary. The KPSS test suggests that the time series is non-stationary. Given that two out of three tests (ADF and Phillips-Perron) indicate stationarity, the time series can be considered stationary, and it is likely that no further differencing is required for time series modeling.

ACF and PACF Plots to Determine Value of p and q: After ensuring stationarity, the Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) plots are analyzed to identify the p and q values. The p value refers to the number of autoregressive terms, while

the q value refers to the number of moving average terms in the ARIMA model. These plots help determine the appropriate lag for both autoregression and the moving average component.

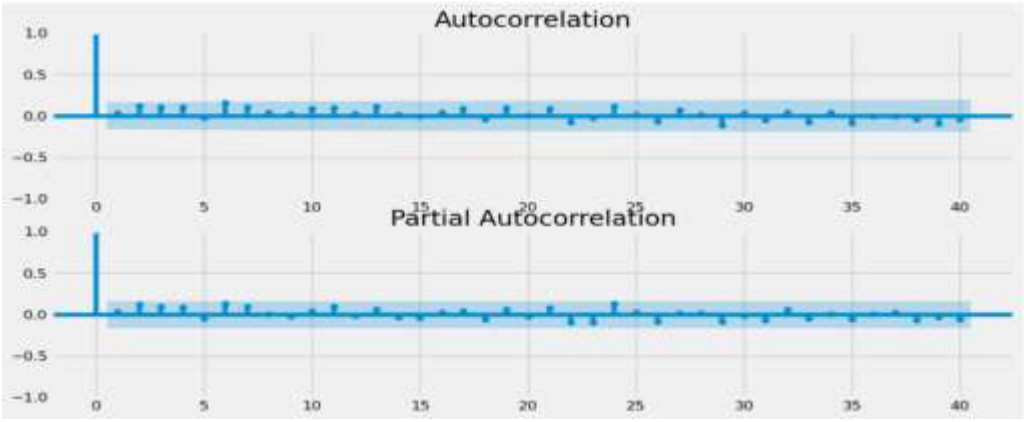


Figure 5: ACF and PACF plots on total records breached

Both the ACF and PACF plots indicate that the time series has short-term dependencies, with only the first lag showing significant correlation. This suggests that a simple ARIMA model with one autoregressive term (AR(1)) and no moving average terms might be a good starting point for modeling this data.

Decide Model: Based on the stationarity test and ACF/PACF plots, the appropriate model is selected. ARIMA is chosen if the data does not exhibit any seasonality, while SARIMA (Seasonal ARIMA) is selected if there is a seasonal pattern in the breach data. The model parameters p, d, q, and any seasonal components (for SARIMA) are determined based on the data characteristics.

ARIMA Models:
ARIMA (AutoRegressive Integrated Moving Average) is a powerful statistical model used for time series forecasting, particularly when the data shows patterns like autocorrelation, trends, or non-stationarity. ARIMA combines three key components: Autoregression (AR), Differencing (I), and Moving Average (MA).
The AR component models the relationship between the current value and its previous values (lags). In an AR(p) model, the current value of the series depends on its previous p values.

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \epsilon_t \quad \text{Eq. 1}$$

where Y_t is the current value, ϕ_1, ϕ_2,…., ϕ_p are the autoregressive coefficients, and ϵ_t is white noise (error term). The differencing component is used to make the time series stationary by subtracting the current value from the previous value. The number of differencing steps is denoted by d. The MA component models the relationship between the current value and past

error terms. In an MA(q) model, the current value depends on the error terms from the previous q time steps.
The complete ARIMA model combines all three components:
p: Number of autoregressive terms.
d: Number of differencing steps to achieve stationarity.
q: Number of moving average terms.
The ARIMA(p, d, q) model is written as:

$$Y_t' = \phi_1 Y_{t-1}' + \phi_2 Y_{t-2}' + \cdots + \phi_p Y_{t-p}' + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

Eq. 2
where Y_t' is the differenced series (after applying the I component).

Fit the Model: The chosen model (ARIMA or SARIMA) is fitted to the historical breach data. During this stage, the model is trained to learn patterns in the data. Techniques like grid search or cross-validation can be used to fine-tune the model's parameters for better accuracy. ARIMA models are widely used because they can handle non-stationary data through differencing, and they capture both the relationships between past values and error terms, making them highly effective for time series forecasting.

VI. Results

Generate Predictions and Forecasts: Once the model is trained, it is used to generate predictions and forecasts for future healthcare data breaches. The model forecasts breach occurrences, highlighting potential risks and vulnerable periods. These predictions are essential for healthcare organizations to pre-emptively enhance security measures and reduce the likelihood of data breaches.

Results of predictions and forecasts by applying various ARIMA models is shown in following figure 6 a) to c).

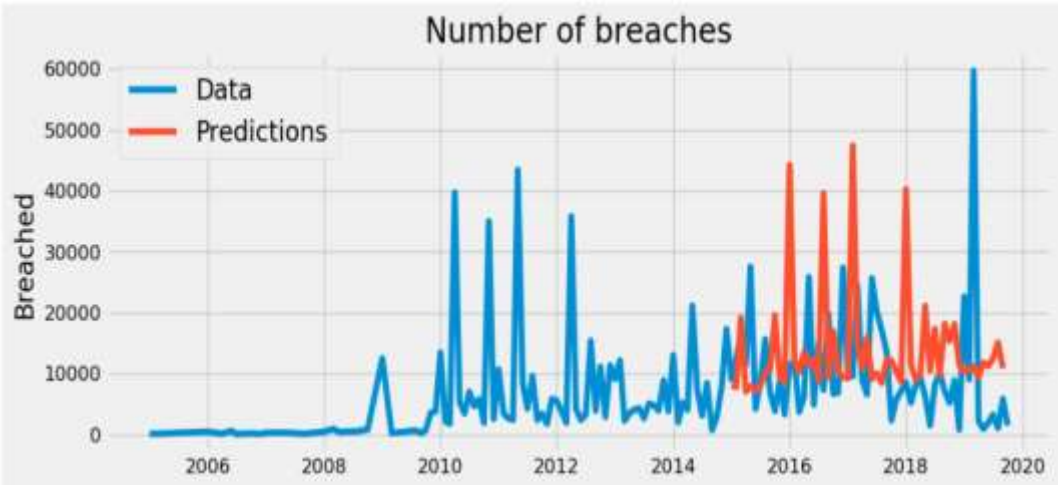


Figure 6 a): SARIMA Predictions on Healthcare Data Breaches

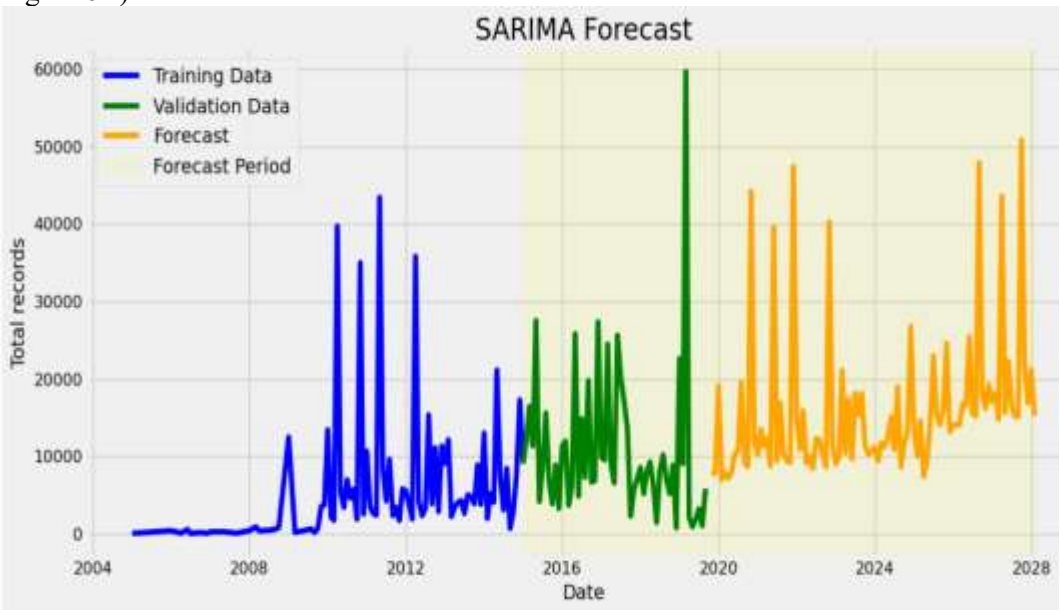


Figure 6 b): SARIMA Forecasts on Healthcare Data Breaches

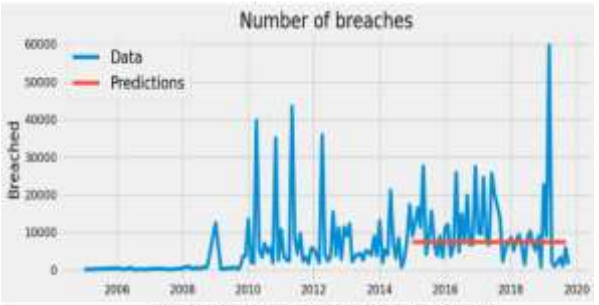
SARIMAX Results

Dep. Variable:	total_records	No. Observations:	86
Model:	SARIMAX(2, 0, 2)x(1, 1, [1], 35)	Log Likelihood	-540.535
Date:	Thu, 17 Oct 2024	AIC	1095.070
Time:	09:24:58	BIC	1108.593
Sample:	0	HQIC	1100.238
	- 86		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1662	0.663	0.251	0.802	-1.134	1.466
ar.L2	0.8310	0.664	1.251	0.211	-0.471	2.132
ma.L1	-0.2166	0.887	-0.244	0.807	-1.956	1.522
ma.L2	-0.6965	0.904	-0.770	0.441	-2.469	1.076
ar.S.L35	-0.9318	0.513	-1.818	0.069	-1.936	0.073
ma.S.L35	-0.0603	1.105	-0.055	0.956	-2.226	2.106
sigma2	1.422e+08	8.63e-09	1.65e+16	0.000	1.42e+08	1.42e+08

Ljung-Box (L1) (Q):	0.12	Jarque-Bera (JB):	111.01
Prob(Q):	0.73	Prob(JB):	0.00
Heteroskedasticity (H):	0.21	Skew:	2.33
Prob(H) (two-sided):	0.00	Kurtosis:	8.52

Figure 6 c): SARIMA Model Summary



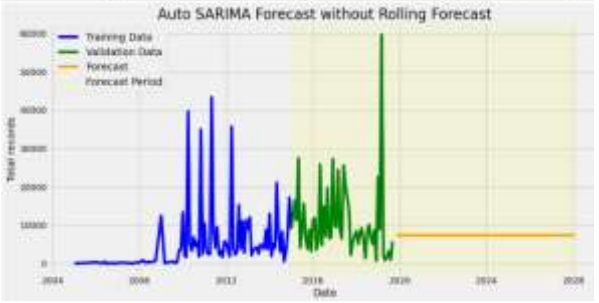
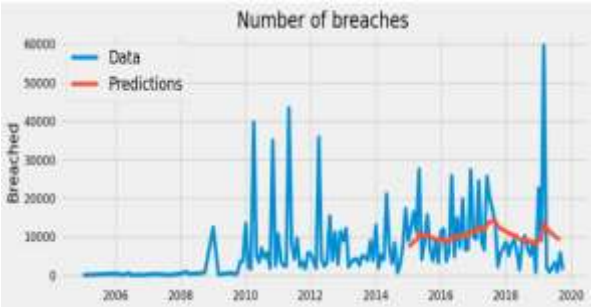


Figure 6 d) Auto SARIMA Predictions

Figure 6 e) Auto SARIMA Forecasts



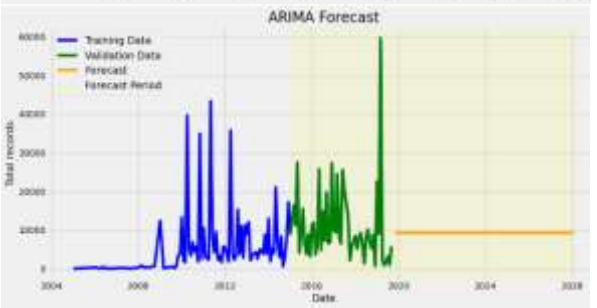
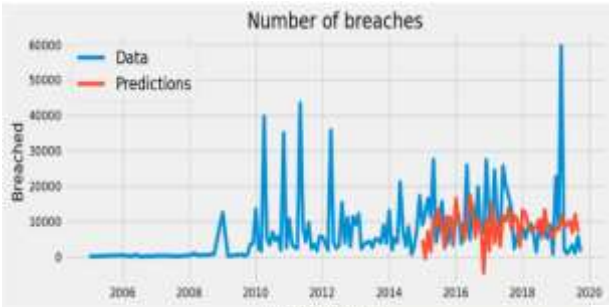


Figure 6 f): ARIMA(2,1,2) Predictions

Figure 6 g): ARIMA(2,1,2) Forecasts



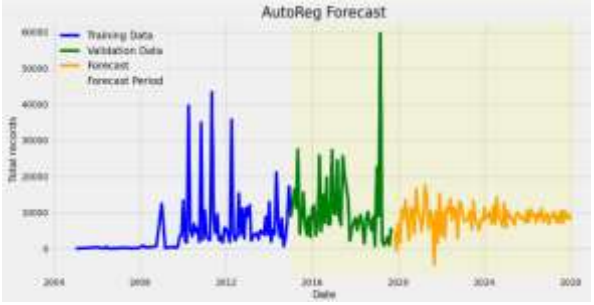
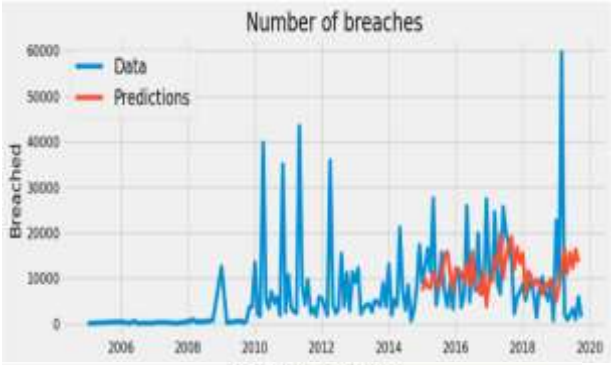


Figure 6 h): Autoregressive (AR) Predictions Figure 6 i): Autoregressive (AR) Forecasts



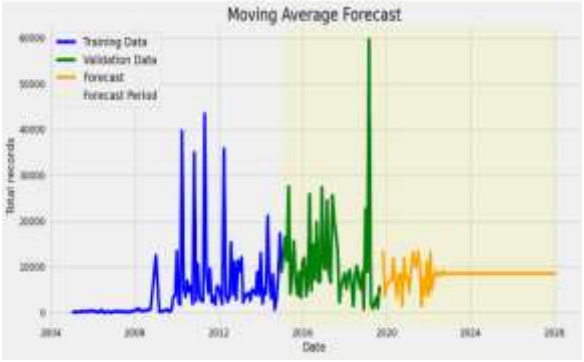


Figure 6 j): Moving Average Model Predictions

Figure 6 k): Moving Average Model Forecasts

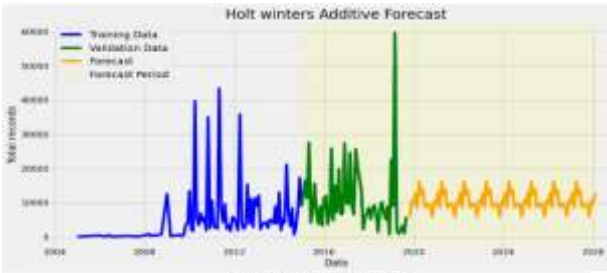


Figure 6 l): HWAF Predictions

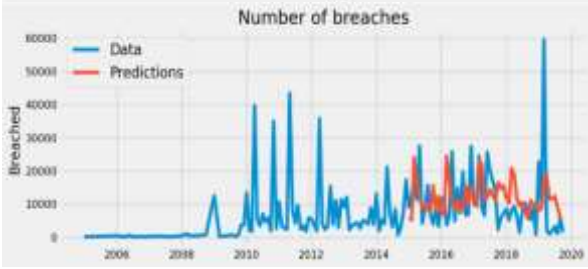


Figure 6 m): HWAF Forecasts

Model	MAE	RMSE	MSE	MAPE (%)
SARIMA	192.04	13,862.12	265,395,284.93	264.79
Auto ARIMA	95.17	10,095.74	56,090,443.49	151.28
ARIMA	97.3	10,088.08	57,924,806.73	154.07
Autoregression	259.13	11,592.62	109,709,262.52	259.13
Moving Average	229.57	11,646.96	107,644,377.79	229.57
Simple Exponential Smoothing	187.78	9,296.51	93,494,156.19	187.78
Holt-Winters Additive	238.78	13,029.54	100,945,917.42	238.78

Table 2: Models performance comparison using various evaluation metrics

Key Insights: As shown in table 2, following key insights are obtained.
Best Performing Model: The Auto ARIMA model exhibited the best predictive performance among all models tested, with the lowest MAE (95.17), RMSE (10,095.74), and MAPE (151.28%). This suggests it is the most reliable model for forecasting healthcare data breaches in this dataset.
Comparative Performance: The ARIMA model closely followed Auto ARIMA, showing similar metrics but slightly worse performance. The Simple Exponential Smoothing model also performed relatively well, with a MAPE of 187.78%, indicating it is suitable for forecasting but not as effective as Auto ARIMA.
Underperforming Models: Both the Autoregression and Moving Average models showed significant prediction errors, with MAPEs of 259.13% and 229.57%, respectively. The Holt-

Winters Additive method demonstrated the poorest performance across all metrics, with the highest MAE (238.78) and MAPE (238.78%).

V. Conclusion

In conclusion, the system architecture for predicting healthcare data breaches leverages time series forecasting models, transforming historical breach data into actionable insights. By following a structured approach from data collection to prediction, healthcare organizations can better manage breach risks, ensuring proactive security measures.

The evaluation of various predictive models for forecasting healthcare data breaches reveals significant insights into their effectiveness. Among the models assessed, the Auto ARIMA model emerged as the most reliable predictor, demonstrating the lowest Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). This suggests that Auto ARIMA can provide more accurate and actionable forecasts, aiding healthcare organizations in proactive breach prevention strategies.

In contrast, models such as Holt-Winters Additive and Autoregression exhibited substantial prediction errors, indicating limitations in their applicability for this specific dataset. The performance of the Simple Exponential Smoothing model was comparatively better but still not as effective as Auto ARIMA. The results underscore the importance of selecting appropriate models tailored to the characteristics of the data and the specific requirements of the forecasting task.

Future Scope

There are several avenues for future research and model enhancement:

Model Refinement: Further optimization of the Auto ARIMA model and exploration of hybrid models could yield improved accuracy. Techniques such as grid search for hyperparameter tuning and incorporating external factors could enhance predictive performance.

Incorporation of Additional Variables: Future studies could benefit from including additional relevant variables, such as organizational size, cybersecurity measures, and historical incident data. This may provide more context and improve forecasting accuracy.

Comparative Analysis with Machine Learning Approaches: Exploring machine learning algorithms such as Random Forest, Gradient Boosting, or Neural Networks may offer alternative solutions that could outperform traditional time series models.

Longitudinal Studies: Conducting longitudinal analyses over extended periods can provide insights into trends and seasonality in healthcare data breaches, further refining predictive capabilities.

Real-time Predictive Analytics: Developing real-time monitoring and predictive analytics tools can assist healthcare organizations in mitigating risks associated with data breaches, providing timely alerts based on model forecasts.

By addressing these areas, future research can contribute to more robust and actionable predictive frameworks for healthcare data breach prevention, ultimately enhancing the security posture of healthcare organizations.

References

[1] Chief Healthcare Executive, "The 11 biggest health data breaches of 2023," Chief Healthcare Executive, 2023. [Online]. Available: https://www.chiefhealthcareexecutive.com/view/these-are-the-11-biggest-health-data-breaches-of-2023
[2] HIPAA Journal, "Healthcare Data Breach Statistics," HIPAA Journal, 2023. [Online]. Available: https://www.hipaajournal.com/healthcare-data-breach-statistics/
[3] IBM Security, "2023 Cost of a Data Breach Report," IBM, 2023. [Online]. Available: https://www.ibm.com/security/data-breach
[4] C. S. Kruse, B. Frederick, T. Jacobson, and D. K. Monticone, "Cybersecurity in healthcare: A systematic review of modern threats and trends," Technology and Health Care, vol. 25, no. 1, pp. 1-10, 2017. DOI: 10.3233/THC-161263.
[5] T. Herath and H. S. B. Herath, "Cybersecurity issues in healthcare information systems: A systematic review," International Journal of Information Management, vol. 51, p. 102059, 2020. DOI: 10.1016/j.ijinfomgt.2019.102059.
[6] S. Verma and D. Dave, "Time series forecasting in cyber risk management: A study using ARIMA models," Journal of Cybersecurity and Privacy, vol. 2, no. 4, pp. 405-426, 2022. DOI: 10.3390/jcp2040022.
[7] N. Kumar and P. Kundu, "Time series big data: A survey on data stream frameworks, analysis and algorithms," Journal of Big Data, vol. 10, 2023. DOI: 10.1186/s40537-023-00651-z.
[8] N. Kshetri, "Ransomware attacks and healthcare breaches: Trends, economic impacts, and policy challenges," Journal of Strategic and International Studies, vol. 18, no. 2, pp. 122-136, 2023.
[9] HIPAA Journal, "Healthcare data breach statistics," 2023. [Online]. Available: https://www.hipaajournal.com/healthcare-data-breach-statistics/
[10] A. H. Lashkari et al., "Time series analysis and prediction for cyber incidents: A data-driven approach," IEEE Transactions on Information Forensics and Security, vol. 14, no. 5, pp. 1283-1297, 2019. DOI: 10.1109/TIFS.2018.2877660.
[11] A. M. Alaa and M. van der Schaar, "Forecasting individual healthcare outcomes using machine learning time series models," PLOS One, vol. 13, no. 3, e0194373, 2018. DOI: 10.1371/journal.pone.0194373.
[12] L. Liu, J. Chen, and W. Li, “Time Series Forecasting of Healthcare Data Breaches Using ARIMA Models,” IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 3, pp. 520-528, 2023.
[13] X. Zhang, Y. Liu, and H. Zhao, “Predicting Healthcare Data Breaches with Ensemble Machine Learning Models,” International Journal of Data Science and Analytics, vol. 10, no. 4, pp. 276-285, 2023.
[14] S. Kumar, R. Agarwal, and P. Sinha, “Ransomware Attacks in Healthcare: A Statistical Analysis of Recent Trends,” Computers & Security, vol. 120, no. 2, pp. 180-190, 2022.
[15] T. Wang, L. Sun, and Q. Yu, “Real-time Monitoring and Prediction of Healthcare Data Breaches: An Integrated System,” Journal of Medical Systems, vol. 47, no. 1, pp. 10-19, 2023.
[16] A. Brown and D. Miller, “Securing Telemedicine: Predictive Models and AI-Driven Cybersecurity,” Health Informatics Journal, vol. 29, no. 2, pp. 135-142, 2023.