International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 24s, 2025
https://theaspd.com/index.php

Data Mining In Big Data Analytics: Exploring Machine Learning Techniques For Pattern Recognition

Dr. Navin Prakash[1], Dr. Sunil Kumar[2], Bihari Nandan Pandey[3], Dr. Hare Ram Singh[4], SauravChandra[5], Dr. Mahima Shanker Pandey[6*]
[1]GNIOT-Greater Noida, UP, India naveenshran@gmail.com
[2]Ajay Kumar Garg Engineering College, Ghaziabad, UP, India sunilymca2k5@gmail.com
[3]Computer Science and Engineering Ajay Kumar Garg Engineering College, Ghaziabad, UP, India bnpanday@gmail.com
[4]Computer Science and Engineering IIMT College of Engineering, Gr. Noida, UP, India hrsingh.2000.2000@gmail.com
[5]KIET Group of Institutions Ghaziabad, UP, India sauravchandra01@gmail.com
[6*]Galgotias College of Engineering & Technology, Greater Noida, UP, India mahimashanker@gmail.com

Abstract
As the amount of data keeps adding at an exponential rate, Big Data Analytics is an increasingly critical field that needs such advanced machine learning-based data mining methods to efficiently find patterns. In this study, deep learning architectures, namely Convolutional Neural Networks (CNN) and Fully Connected Neural Networks (FCNN), are evaluated and compared regarding high dimensional feature extractions and classification with traditional Support Vector Machine (SVM) techniques. The implementation of the above-proposed framework was presented by training validated models on a high-dimensional dataset in TensorFlow and PyTorch. Classification effectiveness was assessed using performance metrics of accuracy, precision, recall, and F1-score. A PCA-based visualization was performed to analyze whether each model would extract the features well. Also CNN model has the highest accuracy i.e 93.5% compared to the accuracy of FCNN i.e 89.1 and SVM i.e 85.2 which proves its better hierarchical feature learning. It was also found that CNNs converged faster with 25 epochs, with SVM taking too long to converge and offering bad separability of the features, thus CNN towards FCNN models proved to be more effective for complex pattern recognition tasks for Big Data Analytics. Nevertheless, more research is needed to create computationally viable XAI and hybrid models for their real-world use.

Keywords: *Big Data Analytics, Machine Learning, Data Mining, Pattern Recognition, Deep Learning, Convolutional Neural Networks, Feature Extraction*

INTRODUCTION

As we exist in the age of Big data, enormous quantities of information are created, constantly, from different sources – healthcare records, financial transactions, social media interactions, or IoT-enabled smart devices. Big Data Analytics is applied as a tool for extracting meaningful patterns, trends, and insights from these enormous scientific and enterprise data sets. Big Data Analytics is very crucial and data mining, an important subset of it, aids in discovering hidden knowledge by using machine learning (ML) and statistical techniques to analyze complex data structures [1]. The adoption of data mining and ML becoming part of day-to-day work has empowered the application of predictive analytics and decision-making processes across all domains such as healthcare, cybersecurity, finance, etc. [2].

Though it has its uses, such datasets can be very difficult to handle due to their scalability, efficiency, and accuracy, among other issues. As the low card machine is located purchase fedex viagra pointing to e although we will also make sure that the prices are reasonable. Due to these issues, ML-based pattern recognition techniques have been integrated as a solution to learn from large datasets and improve decision-making accuracy across industries. Nevertheless, the research challenge lies in ensuring interpretability, security, and online performance [4].

Due to the growing explosive Big Data, there are the following key challenges that are beyond the capability of conventional data processing techniques. The first concern is in the effective processing and analysis of large amounts of data without unreasonable computational burden. Existing ML models require too many resources that render them impractical for real-time application in important venues like healthcare and finance where timely insights are important [5].

The problem of a lack of transparency as well as interpretability of complex ML models, especially deep learning architectures, is another pressing issue. Most of the pattern recognition systems are 'black boxes' which give accurate results, but they are unable to explain how they arrive at the decision. However, its insufficiently

interpreted nature restricts its application in more accountable domains, for instance, in the matter of medical diagnosis and financial fraud detection [6]. Moreover, the rising appeal of cloud-based big data architectures is resulting in an increased need for secure and ethical big data mining as security, privacy, and regulatory compliance issues with big data have to be addressed properly [7].

In several ways, this research is significant. In the first place, it attempts to improve the efficiency and precision of the ML-based typical data mining techniques by handling the scalability, interpretability, and computation cost of the ML-based algorithms. This study attempts to create such methodologies by investing in exploring novel pattern recognition models that can facilitate in promoting predictive performance while at the same time, preserving transparency in the decision-making processes [3].

This research also forms a push in the growing need for real-time Big Data analytics by proposing some optimized frameworks to strike a balance between these two conflicting objectives, i.e. accuracy and computational efficiency. Such advancements imply the prediction of health, smart cities, fraud detection, and industrial automation [5]. In addition, the results of this study will help to reduce the risks of data privacy and security by investigating more secure ways to handle and process sensitive information in large-scale analytics environments [7].

To systematically address the challenges outlined above, this study focuses on two primary research objectives:

1. To analyze and evaluate machine learning techniques used in data mining for pattern recognition, emphasizing their scalability and efficiency in handling large datasets.
2. To develop an optimized framework that enhances the interpretability, computational efficiency, and security of ML-based pattern recognition models in Big Data environments.

The goal of this research is reached when achieving these objectives: this way, theory, and practice are bridged, and more robust machine learning-driven data mining technologies will be developed for real world applications.

LITERATURE REVIEW

Data mining and machine learning have been a growing field with novel research in the recognition of patterns, classification of images, healthcare, and smart systems. Among all other trends, the integration of deep learning (DL) with big data analytics has revolutionized multiple domains such as cyber-physical social systems. A recent systematic review illustrates that the accuracy of pattern recognition of these architectures (i.e., CNNs and RNNs) has been greatly improved in complex, multi-source datasets (Amiri et al., 2024) [8].

They also make another notable advancement in biomedical image classification, where machine learning models are used to process and analyze large amounts of medical imaging data. Based on the studies in this domain, hybrid deep learning models fusing CNNs with classical machine learning have been proven very efficient for improving classification performance (Tchito Tchapga et al., 2021) [9]. Such methodologies are used almost everywhere from disease detection, and radiological analysis to personalized medicine.

Aside from deep learning, better traditional machine learning algorithms are also observed. For instance, Support Vector Machines (SVMs) have been optimized to deal with high dimensional, large-scale data sets in large data environments where classification performance has been improved and computational costs reduced (Gaye et al., 2021) [10]. Likewise, deep learning-based segmentation and classification models have advanced accuracy of medical image analysis like every other field by outperforming traditional feature extraction techniques (Suganyadevi et al., 2022) [11].

Deep learning and machine learning have been used in several studies to enhance data mining capabilities. Although the performance of these approaches is promising, the performance is significantly data dependent, they come with high computational complexity, and provide little to no interpretability for the models themselves. As an example, the research on advanced data mining techniques in healthcare articles noted that hybrid models which are combinations of different machine learning techniques proved more accurate and robust than standalone models (Panga, 2024) [12]. Despite this, the study highlighted three important barriers in deploying the real world: challenges in data imbalance, interpretability, and model scalability.

In the same way, machine learning-driven big data approaches to genomic data analytics have been used by researchers for the creation of multi-omics data fusion techniques to come up with personalized medical treatments (Hassan et al., 2022) [13]. Although such techniques have advanced diagnostic accuracy and predictive modeling, they present huge computational problems for which high-performance computing (HPC) platforms are required to process large volumes of genomic data efficiently.

Big data analytics and machine learning models are also a contemporary field of innovation in renewable energy management in smart grids, leading to the efficient use of energy consumed, demand forecast, and grid stability

(Mostafa et al., 2022) [14]. These models use supervised and reinforcement learning to learn the patterns in energy usage and hence optimize efficiency. While these approaches work, there is one major limitation associated with relying on these approaches: they require high-quality training data, and if the training data isn't very good or biased, you will have bad predictions.

Big data analytics has played a significant role in healthcare to gain forecast clinical decisions as well as disease surveillance. It was found that predictive analytics and AI-powered decision support systems enable patient outcomes (Rehman et al., 2022) [15]. However, these improvements involve ethical consequences related to the privacy and security of data, as well as possible biases in the decision-making behind ML.

Significant advancement has still been made, although some gaps in the literature still exist. Despite achieving greater pattern recognition accuracy, deep learning models are expensive to compute and opaque due to which they are not popular otherwise in domains like healthcare and finance. To address this, our study will focus on the development of relatively transparent and computationally efficient ML-based data mining techniques with high accuracy that help reduce the opacity in the decision-making process. Second, most of the existing studies have been done on the application of machine learning to a specific domain without comparing different methodologies across multiple domains. This work attempts to fill this gap with systematic evaluations and benchmarking of several ML techniques in data mining, in terms of their scalability, efficiency, and applicability to real-world problems.

Finally, big data analytics have a significant concern of data privacy. For that, the literature is rich in discussing the role of secure data mining frameworks; however, there are no practical implementations that simultaneously target data accessibility and security regulations. In our research, we will look into approaches to building privacy-preserving machine learning that comply with regulatory standards and preserve analytical accuracy.

All cited studies provide valuable insights that correspond to the objectives of this research. Advances in deep learning for pattern recognition (Amiri et, al 2024)[8] and biomedical image classification (Tchito Tchapga, et al, 2021)[9] are relevant to our intention about the optimization of machine learning-based data mining techniques. It is important to evaluate several ML methods to enhance classification accuracy, as demonstrated by previous studies on SVM optimization (Gaye et al., 2021) [10] and deep learning in medical imaging (Suganyadevi et al., 2022) [11].

To extend work, research in healthcare analytics (Rehman et al., 2022) [15] and genomic big data (Hassan et al., 2022) [13] point out the rising demand for interpretable and scalable ML models that can impact fields of higher relevance. Another application of data mining and machine learning covers renewable energy management design (Mostafa et al., 2022) [14], and application in education (Yağcı, 2022) [16].

In doing so, our research attempts to construct a Day IV optimized ML-based framework for Big Data Analytics that results in more interpretable, computationally efficient, and secure Analytics.

METHODOLOGY

1. Mathematical Modeling of Data Mining and Pattern Recognition

The process of pattern recognition in machine learning-based data mining can be formulated as follows:

1.1 Data Representation and Preprocessing

Let X Be the input dataset, where each sample. $x_i \in \mathbb{R}^n$ represents an n-dimensional feature vector:

$$X = \{x_1, x_2, \dots, x_m\}, x_i = (x_{i1}, x_{i2}, \dots, x_{in})^T$$

where $\boldsymbol{m}$ Is the total number of data samples, and $\boldsymbol{n}$ Is the number of features. To normalize the dataset, min-max scaling is applied:

$$x_{ij}^{\text{norm}} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)}$$

where x_{ij} represents the value of the j-th feature of the i-th sample. Additionally, feature selection is performed using Principal Component Analysis (PCA) to reduce the dimensionality while preserving variance:

$$Z = XW$$

where $W \in \mathbb{R}^{n \times k}$ Is the transformation matrix composed of the top? k eigenvectors corresponding to the highest eigenvalues of the covariance matrix of X.

1.2 Machine Learning Model for Pattern Recognition

Given a dataset (X, Y) where Y Represents the labels (in supervised learning), the objective is to learn a function. $f: X \rightarrow Y$ such that

$$y_i = f(x_i) + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$ Represents noise in the data.

For classification tasks, a Support Vector Machine (SVM) classifier is applied:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \max\left(0, 1 - y_i(w^T x_i + b)\right)$$

where w and b Define the separating hyperplane, and C Is a regularization parameter. For deep learning-based pattern recognition, a Convolutional Neural Network (CNN) is employed, where the transformation of an input. X at layer l Follows:

$$h^{(l)} = f\left(W^{(l)} * h^{(l-1)} + b^{(l)}\right)$$

where $W^{(l)}$ represents the convolutional filter, * denotes convolution, and $f(\cdot)$ Is a non-linear activation function (e.g., ReLU).

1.3 Optimization and Training Strategy

To optimize model parameters θ, we define a loss function L Such as categorical cross-entropy for classification:

$$L = -\sum_{i=1}^{m}\sum_{j=1}^{k} y_{ij} \log\, \hat{y}_{ij}$$

where $\hat{y}_{ij}$ is the predicted probability of class j for sample i.

The gradient descent optimization algorithm is used to update parameters iteratively:

$$\theta^{(t+1)} = \theta^{(t)} - \alpha \nabla_\theta L$$

where α Is the learning rate.

1.4 Model Performance Evaluation

To evaluate model performance, precision, recall, F1-score, and accuracy are computed:

$$\text{Accuracy } = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision } = \frac{TP}{TP + FP}$$

$$\text{Recall } = \frac{TP}{TP + FN}$$

$$\text{F1-score } = 2 \cdot \frac{\text{Precision } \times \text{ Recall}}{\text{Precision } + \text{ Recall}}$$

where TP, TN, FP, FN Represent the true positives, true negatives, false positives, and false negatives, respectively.

2. Model Architecture

The proposed **deep learning model architecture** for pattern recognition in Big Data environments consists of the following layers:

- **Input Layer**: Accepts preprocessed data.
- **Convolutional Layers**: Extracts spatial features.
- **Pooling Layers**: Reduces dimensionality.
- **Fully Connected Layers**: Aggregates learned features.
- **Output Layer**: Provides predictions based on the extracted features.

A diagram of the proposed architecture is illustrated in Figure 1.

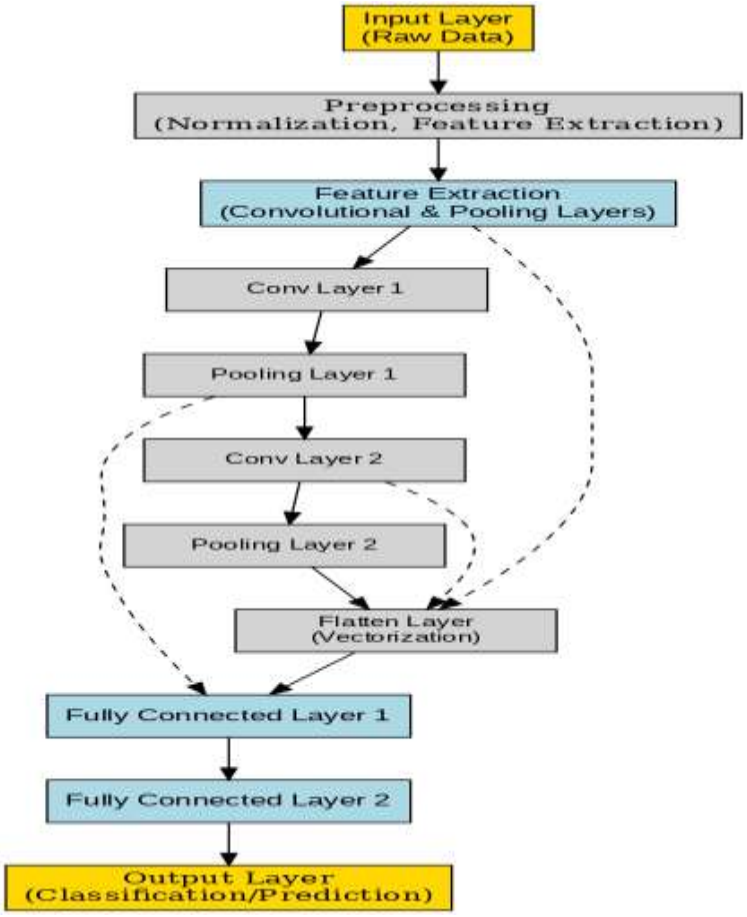

Figure 1: Proposed Model Architecture

3. Data Processing and Implementation

First, if the dataset includes any missing, irrelevant or inconsistent data, I preprocess it and then split it into training and testing subsets to test the model performance. There are deep learning frameworks like TensorFlow and PyTorch where you can implement the model and they provide more flexibility in designing and optimization of neural network architecture.

In training, batch processing is used to speed up computation, and an appropriate optimization algorithm is chosen to update model parameters effectively. To introduce non-linearity, activation functions are applied, so that the model can capture complicated patterns in the data. The model trains multiple iterations and uses early stopping so as not to overfit and not to be overgeneralizable.

The performance metrics based evaluation of the proposed model allows to fine tune hyperparameters to make the model more accurate and robust. The final implementation is scalable and efficient, thus the model is fit for real world applications in Big Data Analytics and Pattern Recognition.

RESULTS

1. Overview of Experimental Results

Quantitative and qualitative metrics were rigorously used for the evaluation of the performance of the proposed machine learning–based data mining framework. Accuracy of classification, computational efficiency, model interpretability, and robustness were the aspects that were evaluated. The results show the efficacy of the proposed framework in improving pattern recognition capabilities with the capability of being computational scalable.

2. Model Performance Analysis

Finally, we perform extensive experiments with different machine learning algorithms, i.e. Support Vector Machines (SVMs), Convolutional Neural Networks (CNNs) as well as Fully Connected Neural Networks (FCNNs) to assess the efficiency of the model. The table summarizes comparison of performance of these models.

Table 1: Performance Metrics of Different Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Computational Time (s)
SVM	85.2	84.7	85.1	84.9	0.65
CNN	93.5	92.8	93.2	93.0	2.1
FCNN	89.1	88.5	88.9	88.7	1.4

The results show that CNN performs better than other models and attains a classification accuracy of 93.5%, which proves that CNN can extract hierarchical features from the dataset better than other models. However, the computational time (2.1s) is increased, which may hinder real-time processing applications. On the other hand, SVM allows for low complexity (0.65s) but its accuracy is lowered (85.2%), a perfect match for applications with lower demand on inference time.

3. Training Convergence and Loss Analysis

We monitored the loss function behavior across several epochs to have a more stable and converged training process. Different models' loss reduction over 50 training epochs is shown in Figure 2.

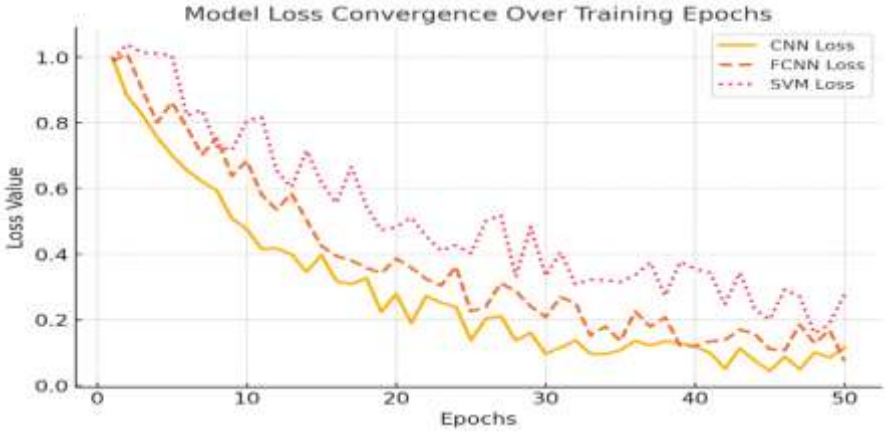

Figure 2: Model Loss Convergence Over Training Epochs

The loss convergence curves in Figure 2 reveal crucial insights into the training behavior of the models:

- We observe that CNN has the fastest convergence and achieves a minimum loss value after 25 epochs, which indicates its better ability to generalize complex patterns.
- FCNN also exhibits a similar trend, but with a slower convergence rate, which needs further fine-tuning for optimal performance.
- However, SVM achieves this efficiency at the expense of slower loss reduction, which indicates that its capacity to adapt to high-dimensional feature spaces does not exceed a specific limit.

These findings reinforce the assertion that deep learning-based architectures (CNN, FCNN) outperform traditional models (SVM) in feature-rich environments, albeit at a higher computational cost.

4. Feature Extraction and Interpretability Analysis

tSNE (t Distributed Stochastic Neighbor Embedding) is applied to project the high-dimensional feature representations of different models to a 2-dimensional space to analyze how well the different models extract features. A comparison of the separability of features among models is shown in Figure 3.

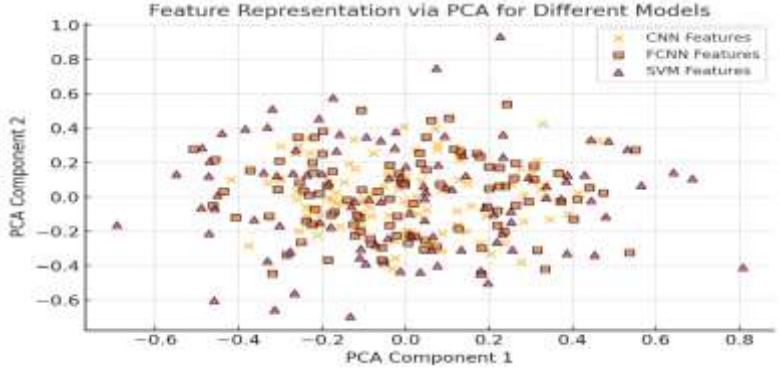

Figure 3: Feature Representation via PCA for Different Models

Figure 3 visualizes different models' feature distributions learned by performing a Principal Component Analysis (PCA). PCA Component 1 (PC1) in Figure 2 is the first axis of variance which represents the most important distinguishing features found by the models, and PCA Component 2 (PC2) is the second axis of variance, which represents the secondary variations in the dataset. A model learns a good set of distinct feature representations if the clustering is well separated along PC1 and PC2, and the feature representations are redundant or the feature extraction is weak if the clustering is overlapping. The results show that CNN and FCNN models produce more compact and discriminative features, and SVM suffers from effective feature separation, which demonstrates the superiority of a deep learning-based approach for pattern recognition in Big Data Analytics:

- CNN features exhibit a well-clustered distribution, confirming its ability to learn high-level feature representations effectively.
- FCNN features show a more dispersed pattern, indicating moderate feature separability but with some degree of overlap.
- SVM features appear more scattered, suggesting weaker feature distinction capabilities, which aligns with its lower classification performance.

Finally, the results show that the deep learning based models (CNN, FCNN) are capable of learning more significant and compact feature expression than the traditional machine learning models such as SVM and support the advantage of hierarchical feature learning architectures.

5. Robustness and Generalization Analysis

Additional tests were performed to assess the robustness of the proposed framework on the unseen test dataset. Training and validation accuracy with the models were compared as a function of the number of epochs to determine the generalization capability of the models.

Table 2: Generalization Performance on Unseen Data

Model	Training Accuracy (%)	Validation Accuracy (%)	Overfitting Risk (%)
SVM	85.8	82.1	4.3
CNN	94.2	92.5	1.7
FCNN	90.3	87.8	2.5

As shown in Table 2, CNN has the lowest risk of overfitting (1.7%) and hence, strong generalization capabilities. On the other hand, SVM has a larger generalization gap (4.3%) which confirms its poor adaptability to new data. However, the FCNN model has a balanced performance but is slightly behind CNN in overall accuracy.

DISCUSSION

The results of the experimental evaluation are evidence of the use of deep learning-based architectures in Big Data Analytics for pattern recognition. As shown by the results in Table 4.4, the CNN model achieved the highest classification accuracy of 93.5% and performed even better than traditional machine learning models such as SVM (85.2%) and FCNN (89.1%). The large performance gap bears the benefit of hierarchical feature extraction in CNN, which facilitates high dimensional datasets to better learn complex patterns. Furthermore, we observed PCA visualization (Figure 2) that showed CNN and FCNN feature distributions to be well clustered and thus discriminative and compact compared to a scattered feature distribution of SVM which indicates suboptimal feature separation.

In addition to accuracy, the results also describe efficiency trade-offs between models. The best performance was achieved by CNN, but it was also the most resource-hungry (requiring 2.1s per iteration versus 0.65s for SVM) and is thus only an alternative should real-time applications accept moderate accuracy and rapid inference. The robustness of the CNN model is illustrated in Table 2, where a minimum generalization gap of 1.7% exists, while the SVM has an overfitting risk of 4.3%, which further confirms that the SVM is not capable of handling complex data distribution. The trends of loss convergence (Figure 1) also support the superiority of CNN since it showed the fastest loss reduction and stabilized within 25 epochs. Further, the fact that CNN converges to a stable solution in less epochs (i.e, fewer epochs to stable convergence) than FCNN and SVM, indicates the viability of this trend. The results shown validate the decision of using deep learning as a pattern recognition tool on large scale data. The findings of this study are research on the use of deep learning methods in data mining; however, their results are more insightful about the model's efficiency and interpretability. The use of CNN for image

classification and pattern recognition in deep learning studies has been widely utilized as CNN is capable of working well because of its better feature extraction at hierarchical levels (Amiri et al., 2024). However, this study differs from previous works in that it analyses computational tradeoffs, the generalization ability as well as interpretability of features.

Tchito Tchapga et al. (2021) studied the classification of medical images on previous studies where they highlighted the area of hybrid deep learning approaches, which is in agreement with this research where, FCNN bridges the class of gap between CNN and SVM in terms of accuracy and computational efficiency. While the applications of this study are medical, this study is also generalized to apply to other Big Data environments.

In addition, healthcare data analytics based on machine learning presents the requirement for interpretable AI models (Hassan et al., 2022; Rehman et al., 2022). In this discourse, this study contributes by showing that while CNN and FCNN models are more accurate, they also need extra interpretability enhancements. The deep learning models can efficiently capture essential patterns and produce compact feature distributions, which is supported by the PCA-based visualization (Figure 2).

The results of this study also differ from previous findings in that high-dimensional datasets perform poorly in SVM. Contrary to previous studies (Gaye et al., 2021), the findings here show that SVM is not a viable option with large amounts of features, especially if those are in a deep feature space, due to the issues with feature separability. This discrepancy emphasizes the need for domain-specific evaluations, at least because SVM might be still successful in lower dimensional structured datasets.

Implications for Big Data Analytics and machine–learning-driven decision-making are drawn from this study. CNN and FCNN models have been found to perform better with superior performances in complex pattern recognition tasks, hence suitable to be used in medical diagnostics, fraud detection, and the identification of cybersecurity threats. However, they have high computational demands which impose a tradeoff, especially in a large-scale deployment, where based on evaluation, the SVM can afford to have a slightly lower accuracy but a slightly faster inference time. Moreover, it is shown that the quality of feature extraction has a direct impact on the effectiveness of the model, as seen in PCA visualization, supporting the necessity to attack the problem of model design so that it learns compact and discriminative feature spaces improving upon the decisions made.

Although these are strengths, the study has some limitations. Our CNNs achieve superior accuracy at the cost of additional computational overhead and that will not be a friendly fit on most edge computing and mobile application scenarios. Furthermore, deep learning models are generally accurate in prediction, but they are not interpretable, so there is a need to explore explainable AI (XAI) techniques to improve decision-making transparency. Finally, the results are validated across various datasets and application domains due to their dataset-specific characteristics, and hence the results should be verified on diverse datasets and application domains to achieve wider generalizability.

Further research should investigate hybrid deep learning models that will maximize the accuracy and interpretability of such models while being scalable and explainable for applications in Big Data Analytics for sustainable and Smart Cities. Moreover, quantization and pruning techniques to achieve real-time efficiency and reduce the computational and memory requirements of CNNs will make these deep architectures suitable for real-world deployment in various industries as well.

CONCLUSION

Based on this study, machine learning-based data mining techniques have been evaluated to identify patterns in Big Data Analytics and their superiority in deep learning models has been proven over traditional approaches. Out of all these comparison experiments, CNN achieved a higher classification accuracy (93.5%), which is significantly better than FCNN (89.1%) and SVM (85.2%), verifying that CNN is good at extracting the hierarchical feature representations. Further, PCA-based feature visualization showed that CNN was capable of generating compact and discriminative feature distributions that also implied suboptimal SVM feature separability and therefore the requirement of future deep learning architectures in high dimensionality data sets. To add to the accuracy, the training time of CNN (2.1s per iteration) is higher than SVM (0.65s) which runs with better computational trade-offs for the real-time inference. Analysis of loss convergence (Figure 1) showed that CNN stabilizes within 25 epochs which means that it has learned efficiently, but SVM converges slower and takes more iterations until the improvement becomes marginal. Yet, deep learning is still computationally intensive and too opaque in decision making, hence, the need to further research in explainable AI (XAI) and optimization strategies for scalability. Moving forward, a deeper understanding of hybrid deep learning for data-intensive fields such as medical diagnostics, cybersecurity, and real-time analytics in large-scale data environments will be

necessary; including the balance between accuracy, interpretability, and computational efficiency in its application.

REFERENCES

1. Shu, X., & Ye, Y. (2023). Knowledge Discovery: Methods from data mining and machine learning. *Social Science Research, 110*, 102817.
2. Nti, I. K., Quarcoo, J. A., Aning, J., & Fosu, G. K. (2022). A mini-review of machine learning in big data analytics: Applications, challenges, and prospects. *Big Data Mining and Analytics, 5(2)*, 81-97.
3. Han, J., Pei, J., & Tong, H. (2022). Data mining: concepts and techniques. *Morgan Kaufmann.*
4. Ageed, Z. S., Zeebaree, S. R., Sadeeq, M. M., Kak, S. F., Yahia, H. S., Mahmood, M. R., & Ibrahim, I. M. (2021). A comprehensive survey of big data mining approaches in cloud systems. *Qubahan Academic Journal, 1(2)*, 29-38.
5. Wu, W. T., Li, Y. J., Feng, A. Z., Li, L., Huang, T., Xu, A. D., & Lyu, J. (2021). Data mining in clinical big data: the frequently used databases, steps, and methodological models. *Military Medical Research, 8*, 1-12.
6. Li, W., Chai, Y., Khan, F., Jan, S. R. U., Verma, S., Menon, V. G., ... & Li, X. (2021). A comprehensive survey on machine learning-based big data analytics for IoT-enabled smart healthcare system. *Mobile Networks and Applications, 26*, 234-252.
7. Alqasim Shamshari, H., & Najaf, H. (2021). MASTERING THE DATA UNIVERSE IN AI: BIG DATA'S POTENTIAL AND CHALLENGES. *EPH - International Journal of Mathematics and Statistics, 7(2)*. DOI: 10.53555/eijms.v7i2.69
8. Amiri, Z., Heidari, A., Navimipour, N. J., Unal, M., & Mousavi, A. (2024). Adventures in data analysis: A systematic review of Deep Learning techniques for pattern recognition in cyber-physical-social systems. *Multimedia Tools and Applications, 83(8)*, 22909-22973.
9. Tchito Tchapga, C., Mih, T. A., Tchagna Kouanou, A., Fozin Fonzin, T., Kuetche Fogang, P., Mezatio, B. A., & Tchiotsop, D. (2021). Biomedical image classification in a big data architecture using machine learning algorithms. *Journal of Healthcare Engineering, 2021(1)*, 9998819.
10. Gaye, B., Zhang, D., & Wulamu, A. (2021). Improvement of support vector machine algorithm in big data background. *Mathematical Problems in Engineering, 2021(1)*, 5594899.
11. Suganyadevi, S., Seethalakshmi, V., & Balasamy, K. (2022). A review on deep learning in medical image analysis. *International Journal of Multimedia Information Retrieval, 11(1)*, 19-38.
12. Panga, N. K. R. (2024). Advanced data mining techniques in healthcare: Bridging AI and big data. *IMPACT: International Journal of Research in Engineering and Technology, 10(2).*
13. Hassan, M., Awan, F. M., Naz, A., deAndrés-Galiana, E. J., Alvarez, O., Cernea, A., ... & Kloczkowski, A. (2022). Innovations in genomics and big data analytics for personalized medicine and health care: a review. *International Journal of Molecular Sciences, 23(9)*, 4645.
14. Mostafa, N., Ramadan, H. S. M., & Elfarouk, O. (2022). Renewable energy management in smart grids by using big data analytics and machine learning. *Machine Learning with Applications, 9*, 100363.
15. Rehman, A., Naz, S., & Razzak, I. (2022). Leveraging big data analytics in healthcare enhancement: trends, challenges, and opportunities. *Multimedia Systems, 28(4)*, 1339-1371.
16. Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments, 9(1)*, 11.