Improved Accuracy of ML Models by Integrating Generative AI and Machine Learning Algorithms for Predictive Analysis

* **Dhruv Singh Negi**, Student, Department of Computer Science, National P.G. College, Lucknow, U.P, India. Email: dhruvsinghnegi27@gmail.com
* **Vatan Shukla**, Student, Department of Computer Science, National P.G. College, Lucknow, U.P, India. Email: shuklavatan05@gmail.com
* **Shweta Sinha**, Assistant Professor, Department of Computer Science, National P.G. College, Lucknow, U.P, India. Email: sinha.shweta020776@gmail.com

Abstract

In recent years, machine learning (ML) has become a crucial tool for predictive analysis across various domains, including healthcare, finance, and customer service. Despite the success of traditional ML models such as Random Forest, Support Vector Machines (SVM), Gradient Boosting, and XGBoost, challenges like data sparsity, manual feature engineering, and the need for domain-specific retraining persist. These limitations often require continuous human intervention, which can impede scalability, especially when adapting to evolving datasets.

This paper explores the integration of Generative AI models, specifically GPT (Generative Pre-trained Transformer), with conventional ML algorithms to address these challenges. By generating synthetic data, these models can enhance training datasets, especially in domains where data is sparse or difficult to obtain. The focus of this work is on text generation, where GPT models are employed to create synthetic text that complements real-world data, improving data quality, balancing class distributions, and reducing the need for extensive labeled datasets.

A practical experiment is conducted using a user behavior phone usage dataset, where synthetic data is generated and feature extraction is performed using Generative AI and combined with real data. The accuracy of models trained on both real and synthetic+real data is tested, demonstrating improved results with the integrated approach. This hybrid model not only enhances predictive accuracy but also reduces the dependency on manual data collection, enabling more scalable and robust machine learning solutions.

Introduction

Machine Learning (ML) has become a cornerstone of modern data-driven decision-making processes across various fields such as healthcare, finance, and customer service. By enabling computers to learn patterns from data and make predictions, ML algorithms have vastly improved the efficiency and accuracy of numerous tasks, including predictive analysis, classification, and anomaly detection. Traditional ML models, such as Random Forest, Support Vector Machines (SVM), Gradient Boosting, and XGBoost, have been widely adopted for their predictive power and flexibility. However, these models are often constrained by several limitations, such as the need for large, high-quality datasets, manual feature extraction, and domain-specific tuning. Additionally, the process of retraining these models to adapt to new tasks or fields remains a time-consuming and resource-intensive endeavor.

[1] Research emphasizes the fusion of LLMs with traditional ML-driven next-generation networks. We argue that although LLMs and ML models show great capabilities, they cannot replace the roles of each other. Combining the generative and reasoning capabilities of LLM with the data analytics capabilities of ML can achieve better results than either model alone can achieve. So we begin by analyzing the capabilities of LLMs and comparing them with traditional ML algorithms. We then explore potential integration by examining the life-cycle of the ML models. Our work also serves as a comprehensive refinement of existing surveys.

In recent years, Generative AI has emerged as a transformative technology, particularly in the domain of text generation. GPT (Generative Pre-trained Transformer) models have revolutionized the way we approach natural language processing tasks by enabling the generation of human-like text. These models can create synthetic data that complements real-world datasets, enriching them and improving overall performance. This capability is especially valuable in situations where obtaining sufficient labeled data is challenging, reducing the reliance on manual data collection and feature engineering.

This paper focuses on the integration of Generative AI, specifically GPT models, with traditional ML algorithms to address the challenges faced by conventional ML systems. By combining synthetic text data generated by GPT models with real-world data, we aim to enhance predictive accuracy, balance class distributions, and reduce the dependence on large datasets. Through an experiment conducted on a user behavior phone usage dataset, we demonstrate the benefits of combining synthetic and real data in improving model performance. Our results show that the hybrid approach yields improved accuracy compared to models trained on real data alone.

The integration of Generative AI with traditional ML models offers a unique opportunity to enhance data diversity, automate feature extraction, and reduce human intervention in the training process. By leveraging the strengths of both approaches, we propose a hybrid model that improves scalability, adaptability, and robustness. [6] Generative Artificial Intelligence refers to artificial intelligence systems with the capability to create text, images, or other forms of media through the utilization of generative models. These models acquire an understanding of patterns and structures within their training data, subsequently generating novel data with akin characteristics. Generative Artificial Intelligence encompasses various types, each tailored for specific tasks or forms of media generation. This paper provides an overview of the theoretical foundations and practical applications of this hybrid approach, highlighting its potential to transform the way predictive models are developed and deployed across various domains.

Background and Literature Review

Generative AI Models
Generative Pre-trained Transformer (GPT) models are a class of language models developed to generate human-like text. Based on the Transformer architecture, GPT models are pre-trained on vast corpora of text data, learning to predict the next word in a sequence. By doing so, they capture a deep understanding of language, syntax, semantics, and context. GPT models, particularly GPT-3 and GPT-4, have become powerful tools for tasks like text generation, summarization, translation, and sentiment analysis due to their impressive ability to create coherent and contextually relevant text across various domains. GPT models excel in Natural Language Processing (NLP) tasks due to their unsupervised pre-training on massive datasets, making them adaptable to a wide range of applications with minimal fine-tuning. This makes GPT particularly useful for generating training data, performing text classification, or automating text-heavy workflows.

Current Applications of GPT in NLP Tasks
GPT has demonstrated success across a broad spectrum of NLP tasks. Some key applications include:

* Summarization: GPT can automatically generate concise summaries of lengthy documents or articles while maintaining essential context and meaning.
* Text Classification: It can be fine-tuned for sentiment analysis, spam detection, and categorization of text data into predefined classes.
* Translation: Pre-trained GPT models have shown proficiency in translating text between languages, reducing the need for manual translation services.
* Text Completion and Dialogue Systems: GPT is the backbone of many conversational AI models, capable of generating responses in chatbots or completing user-initiated queries in natural language.

These capabilities have made GPT indispensable in automating content creation, customer support, and research-related text mining. The model’s pre-training on vast datasets enables it to understand complex topics and generate meaningful, domain-specific text without extensive additional training.

Traditional ML Models
Machine learning algorithms such as Support Vector Machines (SVM), Random Forest, Decision Trees, and XGBoost have been central to many predictive analysis tasks. These models are designed to classify, regress, and make decisions based on large datasets by learning from labeled data. Each model has its own strengths:

* SVM: Useful for high-dimensional data, SVM separates data into classes using hyperplanes, performing well in complex classification tasks.
* Random Forest and Decision Trees: These are tree-based models that work well with tabular data. Random Forest, an ensemble of decision trees, improves accuracy by reducing overfitting.
* XGBoost and Gradient Boosting: These boosting algorithms build multiple weak learners sequentially to create a strong predictive model, excelling in tasks with structured data, particularly in Kaggle competitions and industrial applications.

Challenges with Traditional ML Models
Despite their strengths, traditional ML models face several challenges:

* Data Sparsity: Many ML models struggle with sparse or incomplete datasets, which require significant preprocessing and feature engineering.
* Manual Feature Extraction: For most ML algorithms, the quality of input features significantly impacts model performance. Extracting and selecting the best features from raw data often requires domain expertise and human intervention.
* Domain-Specific Retraining: Most ML models need to be retrained or fine-tuned when applied to new domains or when data distribution shifts, limiting their flexibility across different fields.

Recent Studies on Hybrid Systems Combining AI and ML
Several recent studies have investigated hybrid systems that combine Generative AI and traditional ML algorithms to enhance predictive tasks. Generative AI models, such as GPT and GANs, are being used to generate synthetic data that feeds into traditional ML models, allowing these systems to improve their performance in the absence of large labeled datasets. Studies have shown that GAN-augmented datasets improve the accuracy of image classification models in areas like medical diagnostics and autonomous driving. Similarly, the use of GPT-generated text in NLP tasks has led to significant improvements in text classification models, especially in low-resource settings.

[3] The augmentation methods focus on using diffusion models to enhance image data with flexibility and precision. Randomized Latent Augmentations apply perturbations directly in the latent space, preserving the identity of the original image by using flow matching. The transformation is expressed as [ϕ1]∗([ϕ1]∗(x)+P([ϕ1]∗(x),δ))[ \phi\_1]^\*([\phi\_1]^\*(x) + P([\phi\_1]^\*(x), \delta))[ϕ1​]∗([ϕ1​]∗(x)+P([ϕ1​]∗(x),δ)), where Prand(⋅,δ)P\_{\text{rand}}(\cdot, \delta)Prand​(⋅,δ) introduces Gaussian noise. Modeling Novel Visual Concepts adapts the generative model to work with unseen categories by incorporating new tokens in the text encoder and modifying real images using a splicing technique with noise. This process is defined by xt0=α~t0⋅ϵ+1−α~t0⋅xrefx\_{t\_0} = \tilde{\alpha}\_{t\_0} \cdot \epsilon + \sqrt{1 - \tilde{\alpha}\_{t\_0}} \cdot x\_{\text{ref}}xt0​​=α~t0​​⋅ϵ+1−α~t0​​​⋅xref​, where noise is added at a user-specified time t0t\_0t0​. Object-Centric Augmentations allow selective modification of specific image regions using inpainting and a mask, making the augmentations more intuitive and flexible. Finally, Composing Augmentations combines multiple transformations, creating a hierarchy where augmentations applied at lower time steps (in the latent space) are more abstract, while those at later stages affect the image space. Together, these methods enable diverse, content-aware augmentations that can handle novel classes and fine-tuned modifications.

One promising area of research involves using generative models for feature extraction. For example, by leveraging GPT's ability to understand context, it can automatically generate rich features from text data, which can then be used by traditional ML algorithms. Likewise, GANs can produce detailed visual features that enhance image-based classification tasks.

[4] Synthetic data refers to any data not generated by actual events but artificially created for various purposes, including testing, training, or simulation. Its essence lies in the replication of some properties of the original dataset while not deriving from actual events. On the other hand, generative data is a subset of synthetic data. The distinctive characteristic of generative data is that it is produced using generative models. These models, such as Generative Adversarial Networks or Variational Autoencoders, are trained to understand and capture the underlying data distributions from real datasets. Once trained, they can generate new data samples that are statistically consistent with the original data if it is trained on sufficient data.

However, despite their promise, hybrid systems face challenges in terms of computational complexity and ensuring the quality of generated data. Recent studies have explored these limitations and proposed solutions like transfer learning and semi-supervised training to mitigate them.

Methodology for Integration

Data Augmentation

Generative Pre-trained Transformers (GPT) can play a critical role in data augmentation for Natural Language Processing (NLP) tasks. GPT models are adept at generating high-quality, contextually relevant text, which can help augment training datasets, particularly when real-world labeled data is scarce or imbalanced. By using GPT, synthetic text can be generated to mimic the distribution of the real-world data, enriching the training dataset with varied examples. This technique is especially useful in low-resource settings, where acquiring large labeled datasets is difficult. [7] Synthetic data is not real data, but data that has been generated from real data and that has the same statistical properties as the real data. This means that if an analyst works with a synthetic dataset, they should get analysis results similar to what they would get with real data. The degree to which a synthetic dataset is an accurate proxy for real data is a measure of utility. We refer to the process of generating synthetic data as synthesis.
[2] Generative feature augmentation is explored to synthesize effective training data for few-shot source classes, while effective cross-domain alignment aims to adapt knowledge from source to facilitate the target learning.
For example, if a model is being trained for sentiment analysis on product reviews, GPT can be used to generate additional reviews with varied sentiment and language patterns. This reduces overfitting and improves the generalizability of the ML model. Additionally, GPT can produce paraphrased sentences and diverse textual structures, further enhancing the robustness of the dataset for tasks like text classification, language translation, and summarization.

Cross-Domain Adaptability
One of the key strengths of Generative AI models like GPT is their ability to be fine-tuned for specific tasks or domains with relatively little additional training data. Pre-trained GPT models, which have already learned from vast datasets, can be adapted to new domains by fine-tuning on smaller domain-specific datasets. This allows these generative models to transfer their general knowledge to more specific tasks without requiring the ML pipeline to be built from scratch.

For instance, a pre-trained GPT model could be fine-tuned on financial data to generate high-quality text for tasks like financial sentiment analysis, report generation, or fraud detection. This cross-domain adaptability ensures that ML models built on top of these fine-tuned generative models can generalize better across various tasks, reducing the need for domain-specific retraining or feature engineering.

Data Augmentation and Feature Extraction for User Behavior Prediction: Procedure and Results

To enhance the performance of predictive models on our real-world user behavior dataset, we applied data augmentation techniques, utilizing synthetic data generation to expand the training set. Given the challenges of acquiring large labeled datasets, synthetic data was generated to mimic the distribution of the real-world data, enriching the dataset with varied examples of user behavior. Additionally, we used Generative Pre-trained Transformers (GPT) to automate feature extraction from the unstructured data, further improving the dataset's quality and depth.

Procedure

1. Data Augmentation Strategy: We used a generative model to synthesize additional records, ensuring that the newly generated data reflects the statistical properties of the real data, including App Usage Time, Screen On Time, Battery Drain, Number of Apps Installed, and Data Usage.
2. Synthetic Data Generation: By augmenting the original dataset with synthetic records that preserve the distribution of the real data, we aimed to improve the diversity of the dataset and reduce overfitting.
3. Feature Extraction with GPT: Using GPT’s deep understanding of context and language, we automatically extracted the following 6 key features from the unstructured text data associated with the users (e.g., app descriptions, user feedback, and product reviews):
	* App Usage Patterns: GPT analyzed user behavior logs to capture detailed app usage patterns and time spent per app.
	* Device Preferences: GPT helped extract information regarding user preferences for specific devices or operating systems.
	* Battery Consumption Patterns: By analyzing textual data, GPT was able to infer trends in battery usage linked to certain activities or apps.
	* Data Usage Trends: GPT was used to analyze logs and infer patterns in data usage, especially during specific activities.
	* Age and Gender Associations: GPT generated insights based on the relationship between user demographic factors (age and gender) and their behavior patterns.
	* User Behavior Class: GPT was utilized to identify behavioral clusters or classes based on user habits, which helped in classifying users into different categories (e.g., low, medium, high usage).
4. Training Model: We applied traditional machine learning models, Gradient Boosting, Logistic Regression, Random Forest, Decision Tree, K-Nearest Neighbors and Support Vector Machines to both the original and augmented datasets to evaluate the improvement in prediction accuracy.



Results

The following tables summarizes the accuracy of the models when trained on both the original and augmented datasets, along with feature extraction done by GPT:

Result with Integrated Synthetic and Real Data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 Score |
| Logistic Regression | 0.9619 | 0.9622 | 0.9619 | 0.9619 |
| Random Forest | 1 | 1 | 1 | 1 |
| Decision Tree | 0.9905 | 0.9908 | 0.9905 | 0.9905 |
| K-Nearest Neighbors | 0.9952 | 0.9953 | 0.9952 | 0.9952 |
| Support Vector Machines | 0.9714 | 0.9723 | 0.9714 | 0.9714 |
| Gradient Boosting Machines | 1 | 1 | 1 | 1 |

Results with just Raw Data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 Score |
| Logistic Regression | 0.854 | 0.8552 | 0.854 | 0.8533 |
| Random Forest | 0.9222 | 0.9212 | 0.9222 | 0.9209 |
| Decision Tree | 0.8921 | 0.8935 | 0.8921 | 0.8922 |
| K-Nearest Neighbors | 0.8952 | 0.8945 | 0.8952 | 0.894 |
| Support Vector Machines | 0.8841 | 0.8778 | 0.8841 | 0.8807 |
| Gradient Boosting | 0.9175 | 0.9135 | 0.9175 | 0.9153 |

[Figure 1]Graph that showcases the accuracy across various Metrics along difference ML models in Synthetic+Real Data



[Figure 2]Graph that showcases the accuracy across various Metrics along difference ML models in Real Data



[Figure 3]Graphs showcases the comparison between the Synthetic+Real and Synthetic data results



Advantages of Integration
Enhanced Data Quality
Improved Model Performance through Data Augmentation by Generative AI
The integration of Generative AI models like GPT significantly enhances the quality of data available for training ML algorithms. These generative models can produce high-quality synthetic data that mirrors real-world data distributions, thereby improving the generalizability and accuracy of machine learning models. For instance, GPT can generate text data for low-resource languages or domains. This leads to more diverse, richer datasets, allowing ML models to learn better and produce more reliable predictions. As a result, this integration boosts the overall performance of the models, making them more accurate in handling real-world scenarios.

Handling Class Imbalances in Datasets
One of the critical challenges faced by traditional ML models is class imbalance—a situation where some classes in the dataset are underrepresented. This issue is particularly common in fields like healthcare, where rare diseases may constitute only a small fraction of the data. Generative AI can tackle this issue by generating more examples of the minority class. For instance, GPT can generate text in underrepresented categories, thereby addressing the imbalance. This helps ML models to better understand and predict rare cases, significantly improving their performance in skewed datasets.

Automated Feature Engineering
Reducing Human Intervention in Feature Extraction
In traditional machine learning, feature engineering is a manual, time-consuming process requiring domain expertise. The integration of Generative AI models into the ML workflow automates this task. GPT can extract meaningful, context-aware features from unstructured text data, reducing the need for manual feature extraction. For instance, GPT-generated embeddings can be directly used as input for classification tasks, capturing complex patterns in the text that might otherwise be missed through manual efforts. This automation not only speeds up the model-building process but also improves feature quality, leading to better model performance.

Cross-Field Applicability
Generative AI Models’ Flexibility Allows for Quick Adaptation to Different Domains
One of the most significant advantages of integrating Generative AI with traditional ML algorithms is their cross-domain adaptability. Pre-trained models like GPT can be fine-tuned for different fields, making them highly versatile. For instance, a GPT model trained on a general corpus can be fine-tuned on financial reports to predict stock market trends or on healthcare records for medical diagnostics. This flexibility ensures that ML models built on top of these generative models can be adapted to new domains with minimal retraining, saving time and computational resources.

Improved Interpretability
GPT-Generated Explanations for Model Predictions
Interpretability is a growing concern in machine learning, especially as models become more complex. GPT models can help generate natural language explanations for predictions made by ML models, enhancing their interpretability. For example, in customer service, GPT can be used to generate explanatory text on why a particular response or recommendation was given. In finance, it can explain the factors influencing a stock prediction or a risk assessment, making it easier for users to understand the model's decision-making process. These explanations bridge the gap between the highly technical nature of machine learning models and the need for understandable output for non-expert users.

Robustness and Security
Leveraging Adversarial Examples for Model Robustness
Generative AI models can be instrumental in enhancing the robustness of ML models. By generating adversarial examples—data that is intentionally designed to fool the model—these models can expose weaknesses in ML systems, allowing developers to fortify them against attacks. This is especially important in security-sensitive applications like fraud detection or cybersecurity, where adversaries might try to exploit vulnerabilities in the model. Training ML models with adversarial examples improves their ability to handle unexpected inputs, making them more secure and reliable in real-world scenarios.

Challenges and Limitations
Quality Control of Generated Data
Ensuring Representativeness and Minimizing Bias in Synthetic Data
One of the critical challenges in integrating Generative AI models like GPT with traditional ML workflows is maintaining quality control over the synthetic data these models generate. Synthetic data, although useful for augmenting training datasets, may not always fully capture the intricacies of real-world data. There is also the risk that biases inherent in the training data could be amplified in the generated data, leading to inaccurate or skewed predictions when fed into ML models.

To address this, rigorous validation mechanisms must be implemented. Synthetic data should be evaluated against real-world data to ensure it is representative and unbiased. Regular checks for data quality, diversity, and distribution are necessary to prevent unintended consequences such as model overfitting, poor generalization, or biased outcomes. Employing techniques like adversarial training, regularization, and post-generation auditing can help mitigate the risks of using synthetic data.

Model Complexity
Increased Complexity from Integration
The integration of Generative AI models with traditional ML algorithms introduces significant complexity into the overall system. Generative models like GPT are inherently deep and computationally intensive due to their large architectures. When combined with traditional ML algorithms, this adds layers of complexity in terms of model architecture, training processes, and optimization requirements.

For example, in a system where GPT is used for text generation and feature extraction while a Random Forest is used for classification, there is a need to fine-tune the interaction between these models to ensure that the outputs from one are effectively fed into the other. This can be challenging as each model may have different requirements for data preprocessing, optimization, and tuning. Additionally, hyperparameter tuning for both generative models and traditional ML models can become a multi-dimensional problem, requiring more sophisticated optimization strategies.

This increase in complexity also leads to the need for more computational resources, making it essential to have optimized pipelines, efficient memory management, and scalable architectures to ensure smooth integration without incurring performance bottlenecks.

Future Directions and Research Opportunities

Automating the ML Pipeline

The integration of Generative AI into machine learning systems opens the door to the full automation of the machine learning pipeline. This includes tasks like data preprocessing, feature extraction, and model tuning, which are currently labor-intensive processes requiring significant human intervention. For instance, GPT models could be used to automatically clean, process, and augment textual datasets, while GANs can enhance image datasets by generating synthetic images that expand the range of training data. Moreover, automated hyperparameter tuning using advanced generative algorithms could allow models to adjust their own parameters for optimal performance, reducing the need for manual trial-and-error methods. This would not only make the machine learning process faster and more efficient but also more accessible to non-experts.

Future research could focus on refining these techniques to further reduce human oversight in ML systems. This includes exploring meta-learning (where models learn how to optimize themselves) and advancing the use of Generative AI for real-time feedback loops that adapt and retrain models dynamically based on new data inputs.

Conclusion

The integration of Generative AI models, such as GPT (for text) and GANs (for images), with traditional Machine Learning (ML) algorithms represents a transformative approach to predictive analysis. By leveraging the strengths of generative models in data augmentation and feature extraction, we can significantly enhance the quality of data, which directly improves the performance of ML models. This synergy addresses longstanding challenges such as data scarcity, class imbalances, and the need for manual feature engineering, making the machine learning pipeline more efficient and reducing human intervention.

Moreover, this hybrid approach allows for increased cross-domain adaptability, where models can be fine-tuned to handle a wide variety of tasks without frequent retraining. Whether it’s in healthcare, where GANs can generate synthetic medical images to improve diagnostic accuracy, or in finance, where GPT can provide natural language summaries and explanations of financial predictions, the integration of generative AI and ML holds immense potential. The ability of generative models to offer both interpretability and robustness—such as by generating adversarial examples to strengthen security—also addresses concerns around model trustworthiness.

Looking forward, the combination of Generative AI with ML algorithms is poised to make a significant impact across industries like healthcare, finance, and e-commerce, where data-driven predictions are critical. This approach not only improves the scalability and flexibility of predictive models but also enables more transparent and ethical AI systems. The ongoing research into automating the machine learning pipeline and addressing bias will further refine these integrated systems, pushing the boundaries of what predictive analytics can achieve in various real-world applications.

References

[1] Han Zhang, Akram Bin Sediq, Ali Afana, and Melike Erol-Kantarci, Senior Member, IEEE, Generative AI-in-the-loop: Integrating LLMs and GPTs into the Next Generation Networks

[2] T. Wang, Z. Ding, W. Shao, H. Tang and K. Huang, "Towards Fair Cross-Domain Adaptation via Generative Learning," 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2021, pp. 454-463, doi: 10.1109/WACV48630.2021.00050

[3] Shivashankar Shane Miller , Semantic Data Augmentation with Generative Models

[4] Yunhao Chen; Zihui Yan; Yunjie Zhu , A Comprehensive Survey for Generative Data Augmentation

[5] Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers. Staphord Bengesi, Hoda El-Sayed, Md Kamruzzaman Sarker, Yao Houkpati1, John Irungu and Timothy Oladunni

[6] Generative artificial intelligence: a systematic review and applications Sandeep Singh Sengar ·Affan Bin Hasan·Sanjay Kumar·Fiona Carroll

[7] Khaled El Emam, Lucy Mosquera & Richard Hoptroff, Practical Synthetic Data Generation Balancing Privacy and the Broad Availability of Data