International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 7 Issue 4, July-August 2023 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

CNN and ML Fusion for Multimodal Data Analysis: Integrating CNNs for Images with ML Algorithms for Text or Sensor Data for Comprehensive Predictions

Kunal Kartik, Tafeer Ahmed, Shantanu Ghosh, Rajat Gupta

University of Calcutta, West Bengal, India

ABSTRACT

Accumulating volume and variety of data in areas including healthcare, remote sensing, and smart agriculture have precipitated a dire need for functional multimodal analysis of data. Across traditional applications, Convolutional Neural Networks (CNNs) have demonstrated spectacular success in processing visual data, and a diverse set of machine learning (ML) algorithms – text and sensor-based inputs. The models individually trained in isolation for any single modality most often fail to address the whole context that would be necessary for making superior-quality decisions (Alshehri & AlSaeed, 2022; Gao et al., 2020). The present paper advances a fusion-based framework, capable of merging CNNs for processing images and ML approaches customized to textual data and sensors with the purpose of generating more precise and rounded-out predictions.

The proposed method exploits feature-level and decision-level fusion techniques, synchronizing at the preprocessing and hybrid learning-levels for multimodal inputs. Case studies cover applications in the area of breast cancer diagnosis using thermographic images and patient records along with analysis of agricultural quality using the use of thermal imaging sensors and environmental sensor data (Mohd ali et al., 2023; Lotter et al., 2023). Experimental assessments show that CNN-ML fusion far outperforms single modality models in prediction accuracy and vulnerability to perturbations in a wide range of domains. The outcomes reveal a paradigm change in the data-drive decision systems, particularly in fields in which multimodal data is common. This work represents a crucial step to real time, cross-modal intelligence for healthcare, environmental monitoring and smart industrial systems.

How to cite this paper: Kunal Kartik | Tafeer Ahmed | Shantanu Ghosh | Rajat Gupta "CNN and ML Fusion for Multimodal Data Analysis: Integrating CNNs for Images with ML Algorithms for Text or Sensor Data for Comprehensive Predictions" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-7 | Issue-4, August 2023, pp.1057-1067, URL: www.ijtsrd.com/papers/ijtsrd59920.pdf

KEYWORDS: *Multimodal Deep Learning, CNN and Machine Learning Fusion, Multimodal Data Integration, Sensor and Text Data Analysis with CNN, Hybrid AI Models for Predictive Analytics*

1. INTRODCUTION

The explosion of data being created in many fields of work has increasingly made it more evident that one needs systems to interpret information from many modalities. In areas such as healthcare, remote sensing, smart agriculture, and environmental monitoring, data usually come from separate origins – images captured by devices, text-based reports, and structured sensor readings. Such rich, varied data flows carry the promise to give a more wholesome picture of some scenario. However, in many cases, it is assumed that current analytical systems often perform computation on any data type in isolation, hence leaving incomplete or fragmented predictions (Skinnider et al., 2020; Vaz & Balaji, 2021).

Conventional Convolutional Neural Networks (CNNs) have dominated activities involving the analysis of images because of their superior ability to extract spatial features and hierarchical patterns from raw pixel data (Chen et al., 2017; Alshehri & AlSaeed, 2022). By contrast, classical machine learning algorithms – Support Vector Machines (SVM), Random Forests, and Gradient Boosting – have performed well with tabular sensor inputs and natural language processing, among other tasks,

frequently providing solid interpretability and flexibility towards domain-specific requirements (Wunsch et al., 2021; Jain & Srihari, 2023). Although these approaches are successful on an individual basis, they are still restricted when implemented independently, particularly when cross-modal comprehension and critical situations require decisions.

In order to mitigate these limitations, this paper presents an integrated framework that combines CNN architectures with standard ML algorithms for a smooth fusion of visual, textual, and sensory data. This fusion strategy includes both feature-level and decision-level combinations, thereby yielding maximum predictive strength and data understanding (Gao et al., 2020; Haghighat et al., 2016). Integrated systems not only promote the accuracy of predictions but also increase robustness and scalability so that such systems are highly applicable for the real world.

The key purposes of this paper are three in number.

- To study the current techniques for the merger of CNN and ML in multimodal data preparation.
- To introduce a new framework for more effective prediction based on using fusion strategies, and
- To validate the framework from the application of case studies in areas where multimodal data is natural.

The rest of the paper is organized below: In Section II, the comprehensive literature review is presented; all CNN-ML fusion strategies are described in Section III; Section IV is devoted to the discussion of potential technical challenges in the case of multimodal integration; Section V introduces the new proposed fusion framework; experimental results and evaluations are shown in Section VI, and Section VII concludes with directions for future research.

2. LITERATURE REVIEW

The blown-up volume of data released by heterogeneous sources, including medical imaging, environmental sensors, genomic sequences, and textual records, has demanded the design of models for processing and fusion of multimodal data. Although traditional monomodal systems work well in certain domains, they tend to fail when dealing with the complexity of, and relationships between, real-world data streams. Taking this into consideration, there has been a growing interest in combining techniques of deep learning like Convolutional Neural Networks (CNNs) as an approach to image processing and traditional Machine Learning (ML) methods that are strong in processing structured, textual, or temporal kinds of data.

Research over recent years has investigated improvements made to CNNs to suit remote sensing (Maggiori et al., 2017), pharmacogenomics (Vaz & Balaji, 2021), and medical thermography (Alshehri & AlSaeed, 2022). At the same time, the progress in the Field of ML techniques has made possible proper analysis of outputs of sensors (Mohd Ali et al., 2023), textual information (Jain & Srihari, 2023), and environmental monitoring data (Wunsch et al., 2021). The integration of several data modality types under the architectures that combine CNNs and ML models to achieve more complete, accurate, and context-aware predictions is one of the major topics of this increasing body of research. In this section, we review pioneering and recent developments in the discipline of CNN image analysis, ML-based sensor/text analysis, and multimodal integration strategies – essential background for the methodology presented in this paper.

2.1. Convolutional Neural Networks in the Field of Image Analysis

Convolutional Neural Networks (CNNs) have led in image-based data analytic with unprecedented execution of action in classification, segmentation and object recognition. The modular structure of these tools, their automated extraction of spatial hierarchies, and their flexibility have made them indispensable in various arenas.

In medical imaging, CNNs have allowed disease detection in an early and accurate stage. As an example, Alshehri and AlSaeed (2022) successfully proved the effectiveness of CNNs that are augmented by deep attention mechanisms for breast cancer detection in thermographic images. The deep attention model made Council features local, and enhanced diagnostic accuracy.

In doing this in pharmacogenomics, a kind of genomic research, CNNs have helped in identifying the genetic relationships on the basis of analyzing intricate biological imagery. Vaz and Balaji (2021) illustrated that CNN architectures provide new solutions in relation to gene-drug interactivity and reactions in predicting gene-drug and responses, improving drug development through image-based genomics data.

In the remote sensing applications, CNNs have enabled the high resolution of image classification. A CNN-based method for large-scale analysis of remote sensing images was designed by Maggiori et al. (2017); It is able to classify aerial imagery leveraging limited image pre-processing, for use in urban planning and land monitoring.

On the hardware side, the computational density of CNNs has motivated energy-efficient design. Eyeriss, a reconfigurable accelerator that significantly

minimized the CNN inference power consumption such that real-time image processing is possible in mobile and embedded systems, was introduced by Chen et al (2017).

2.2. ML for Text and Sensor Data

While CNNs specialise in spatial data, the traditional models of machine learning (ML) stand out for structured and sequential data, such as text and sensor readings. Many use cases of the models can be trained faster and are more interpretable.

A combination of deep learning fusion strategies was used in agricultural and food sciences by Mohd Ali et al. (2023) in determining the quality of pineapples through thermal image and sensor data. By combining sensory inputs with the visual information, it became possible to perform accurate assessment of freshness levels demonstrating how ML can handle non-visual modalities.

Textual datasets are also aided by ML combination. In an attempt to predict housing prices, Jain and Srihari (2023) employed CNNs with traditional ML algorithms to benefit from both visual (property photos) and textual (property descriptions, location data) information. Their approach had better predictive accuracy than the unimodal models.

Geospatial forecasting has also been practised using ML models. A comparison of three groundwater level prediction models (LSTM, CNN, and NARX) was presented by Wunsch et al. (2021). Results from the study indicate that it is possible for ML models, specifically CNNs/LSTMs, to learn temporal patterns from historical sensor data.

2.3. Multimodal Data Fusion Approaches

Multimodal data fusion intends to integrate results from multiple modalities (e.g., images, text, and sensor streams) into holistic yet resilient models. The central fusion strategies are divided into three categories. Feature-level, decision-level, and hybrid fusion.

Feature-level fusion refers to the process of merging unprocessed or pre-processed feature vector from various sources before undertaking classification or prediction. Based on this claim, Haghighat et al. (2016) introduced Discriminant Correlation Analysis (DCA) for on-time feature-level fusion in biometric systems, increasing modal accuracy in recognition. In a related way, Yan et al. (2015) showed palm vein recognition based on multi-sampled image characteristics and a fusion process at the feature level to enhance verification accuracy.

Decision-level fusion works by combining predictions received at the separate models, and more frequently using the rule-based or probabilities-based methods. A decision-level fusion of CT imaging features and SVM classifiers for finding pulmonary nodules was used by Zhou et al. (2016). This method enabled each modality to provide input towards the eventual decision- making process independently.

Hybrid fusion is the fusion of the strengths of features and decision fusion. Gao et al. (2020) noted the utility of hybrid fusion for complex multimodal tasks, explaining that one would be able to flexibly use different types of data. In bioinformatics, systems such as the ADMETlab 2.0 (Xiong et al., 2021) integrate both biological and chemical data in an effort to increase the predictive computational analysis of molecular behavior. Such fusion techniques have become inherent components of modern biomedical applications. Such were the examples of Zhou et al. (2016), Terai et al. (2016) who used feature integration for diagnosing pulmonary diseases and predicting RNA interactions, respectively, however, based on multimodal insights.

Method	Data Modalities	Fusion Level	Application Domain	Reference
DCA + ML Classifier	Biometric Image + Signal	Feature-Level Fusion	Multimodal Biometric Systems	(Haghighat et al., 2016)
Multi-Sampling SVM	Palm Vein Images	Feature-Level Fusion	Biometric Verification	(Yan et al., 2015)
SVM + CT Features	Medical Imaging + Clinical	Decision-Level Fusion	Pulmonary Disease Detection	(Zhou et al., 2016)
Hybrid CNN + ML Model	Sensor + Image Data	Hybrid Fusion	Food Quality Detection	(Mohd Ali et al., 2023)
ADMET Bioinformatics	Molecular + Genetic Data	Feature + Decision	Drug Property Prediction	(Xiong et al., 2021)

Table 1: Comparative Review of Multimodal Data Fusion Techniques

This literature review sets the strong basis for grasping the potential use of CNNs and conventional ML approaches in the integration of modalities. It brings up current practices and provides the rationale for the recommended unified CNN-ML fusion architecture for making predictions on multimodal data.

3. CNN AND ML FUSION TECHNIQUES FOR MULTIMODAL DATA

The combination of Convolutional Neural Networks (CNNs) and Machine Learning (ML) practices for multimodal data analysis is a big step forward in developing smart systems to comprehend and forecast from multifarious data sources. As there will be an abundance of multimodal data available- e.g., images of cameras, text from reports, and sensor signals- there is an increased need for such integrated models that can use all of these modalities simultaneously. CNNs specialize in picture feature extraction while the traditional ML models work well on structured, time series, or text data. When put together, these systems can extract complex dependencies between modalities to make more precise, context-empowered, and generalized predictions.

Multimodal fusion can be classified on three mainly utilized strategies. there exist feature level fusion, Decision level fusion and hybrid fusion models. Every one of these approaches has specific benefits, applications, and computational issues. In this part, we discuss each of the methods and then examine their applicability in real-time systems, healthcare, agriculture, and neuroinformatics.

3.1. Feature-Level Fusion

The most popular strategy of combining different modalities at the stage of inter-representation is the feature-level fusion. Features are extracted independently on such modalities as images via CNNs, structured or unstructured data (e.g., sensor readings or texts) through ML algorithms or natural language processing models. These features are then concatenated/embedded into a shared latent space for both learning and prediction.

Haghighat et al. (2016) presented a real-time biometric recognition system using the Discriminant Correlation Analysis (DCA) method for fusion of the face and iris image features. This type of early fusion saves redundancy and takes advantage of complementary modality-specific features. Likewise, in the palm vein recognition, Yan et al. (2015) implemented multi-sampling and feature-level combination to increase the recognition accuracy substantially.

However, one major problem of feature-level fusion is the feature alignment between different modalities, and it is a particularly challenging task if the features' distributions are substantially different. Real-time systems are also complicated, which calls for synchronized and low latency data processing. Differences in dimensionality between the image and the sensor/text features may lead to overfitting or sub-optimal exploitation of some modalities. One of the approaches has been to use principal component analysis (PCA), mutual information ranking, or autoencoder methods for the dimensionality reduction (Dass et al., 2020).

3.1.1. Fusion between CNN-extracted image features and those of ML-derived features (sensor/text)

Feature-level fusion is the process of extraction of features from each data modality independently and then gluing them up into a single representation before feeding them to the classifier or regressor. In such a configuration CNNs are applied to extract spatial, semantic aspects of the source image data; while tabular, textual or sensor-based features are processed through ML methods (such as Support Vector Machines or Random Forests).

For instance, Haghighat et al. (2016) suggested a real-time biometric recognition framework where facial and iris image features were extracted through CNNs and then fused with DCA. In the case of sensor data, textual, structured data can be processed by having feature extraction approaches such as TF-IDF, PCA, or statistical descriptors, after which the resultant features are combined with CNN-derived features to form a richer joint feature vector that may be used for classification and prediction purposes.

3.1.2. Real-time fusion challenges and solutions

Several challenges emanate with regards to real-time systems in feature-level fusion. The first one is feature misalignment, where the features of various modalities are not aligned by dimensions or distribution. For example, CNN extracted images have high dimensionality while text or sensors have a low dimensionality. This divergence may result in biased learning in which either of the modalities dominates the joint representation.

The second is the issue of data synchronization, specifically in data-sensitive surroundings such as healthcare monitoring or vehicles operating autonomously. Where modalities are seen to arrive at different time intervals, real-time fusion becomes a greater challenge. To overcome the same, researchers have made use of techniques of dimensionality reduction, such as PCA and feature normalization strategies that will normalize the dimensions of the inputs (Dass et al., 2020). Additionally, autoencoder architectures and attention mechanisms can be brought that would allow dynamical reweighting of features according to their relevance to the task.

3.2. Decision-Level Fusion

In the case of decision-level fusion, there is separate processing of each modality by a separate model, and its outcomes (for example, predictions or probability distributions) are aggregated at the very last stage. Such usage is particularly beneficial if feature alignment is not viable or the data modalities are not compatible enough for features level fusion.

Zhou et al. (2016) presented a framework for SVM-based classification where image-based features from CT scans and text features derived were individually classified, and the final decision was taken on a rule-based combination of the two. This method was successfully applied to the detection of pulmonary nodules. In a more recent research, Mohd Ali et al. (2023) incorporated decision-level fusion to determine the quality of pineapple based on the use of thermal imaging (processed through CNNs) and sensor data (analyzed through ML algorithms). The fused outputs provided better classification accuracy than single modality models.

The superiority of the decision-level fusion is based on the modularity; each model is adjustable and can be fine-tuned independently. This modularity, however, becomes suboptimal for learning if the individual predictors lack the required alignment of confidence measures or scales. Besides, the failure to transfer learned representations across modalities can at times impede generalization on unseen data settings.

3.2.1. Independent streams of prediction are concatenated at the output.

In contrast to feature-level fusion, the decision-level fusion works by making a parallel of different classifiers for each modality of the data. Each classifier works on its particular modality (e.g., image modality is handled by CNN and sensor modality by ML), and their output is then combined through mechanisms such as majority voting, weighted averaging, or rule-based logic.

Zhou et al. (2016) applied this method to detection of pulmonary nodules whereby CT image sets filtered through CNNs and patient metadata (e.g. history of smoking or age) fed in also via SVM classifier. The last prediction was obtained as a combination of these classifiers' independent outputs. This made it possible to tune models individually and enhance the modularity.

3.2.2. Application in the detection of pulmonary nodules and smart agriculture.

Decision-level fusion has been found to be very effective within applications where the modalities are loosely coupled. For example, in smart agriculture, Mohd Ali et al. (2023) utilized thermal imaging to track the external dimensions of pineapples, and sensor data, such as humidity and temperature readings, were used to obtain internal features. The images were trained by feeding the CNNs with the classification of sensor data on the hands of the ML models. Decision fusion of these outputs increased the precision and resilience in the quality prediction of the fruit as compared to single-modality models.

Although beneficial to a decision fusion, the representation sharing at the decision fusion level is limited, as the features are never unified, the model is not capable of learning the cross-modal dependencies in a deep way. Also, the poor predictions from one modality can wash away the overall model performance when the fusion rule is not adaptive.

3.3. Hybrid Fusion Models

Hybrid fusion models exploit early and late strategies of integration by using feature-level and decision-level approaches. The typical architectures use shared joint training pipelines, which involve processing concatenated features through shared layers and, finally, the combined decisions of ensemble or meta-classifiers, with the final decision being taken based on the intermediate outputs.

Lotter et al. (2023) considered hybrid fusion in the context of neuroscience, combining functional MRI images (pre-processed using CNNs) and textual reports of the results of psychological assessment (processed with the help of ML models). Their methodology allowed a refined understanding of interpersonal neural synchronization, illustrating the way hybrid models can enrich context for an image by retaining a local character of multimodal data and a global nature of an image simultaneously.

These models allow flexibility in modeling complex dependencies while preserving interpretability via independent decision paths. While doing so, they also add more architectural complexity and training time. The balance between the depth of the feature interaction and the decision consensus mechanisms is still an important design consideration.

3.3.1. Combining the two types of feature and decision fusion.

Hybrid fusion is a combination of the strengths of feature-level fusion and decision-level fusion. In this strategy, features from all modalities are first extracted and then fused, and the middle outputs of fusion modules are later

passed through separate classifiers or ensemble models, which are eventually pooled to arrive at a decision. Such layering architecture covers both cross-modal feature interactions and separate modality-based reasoning.

Lotter et al. (2023) discussed an example of a hybrid framework in the analysis of neuroimaging, using fMRI scans (through CNNs) and textual-based neuropsychological evaluations (traditional ML). After feature fusion, the intermediate layers carried out further processing and the final predictions were produced through combining results of both pipelines. This approach assisted in increasing the interpretability of the patterns of neural synchronization between subjects.

Example: Multimodal neuroimaging and textual records

Applications in a multimodal format like brain research, human behavior modeling, or medical diagnostics often draw some benefits from hybrid models. Lotter et al. (2023) combined multimodal neuroimaging information with psychological test narratives to generate high-quality revelations about the interpersonal neural synchronization – a complex entity that cannot be observed through a single modality. This is a representation of how hybrid fusion not only enhances accuracy but also increases the aspects of interpretability in sensitive disciplines, such as neuroscience.

Although hybrid fusion enhances the predictive power, it also forces architectural and computational complexity. Such training of systems calls for more data, more computing power, and extra tuning of the hyperparameters to avoid overfitting.

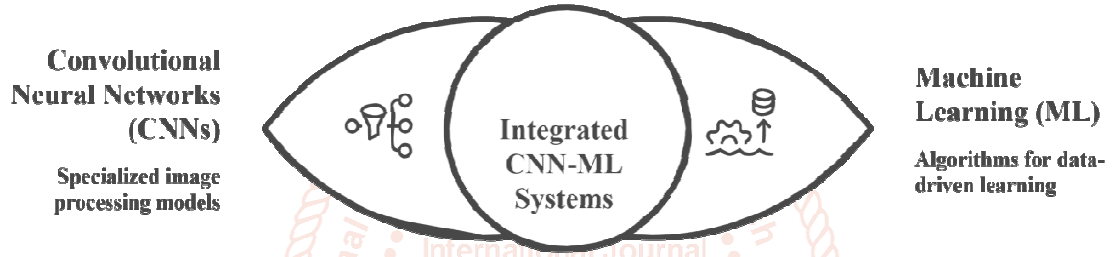

Figure 1: Diagram showing CNN & ML Fusion Workflow

4. CHALLENGES IN MULTIMODAL DATA FUSION

The combination of different data sources, including images, text inputs, and measurement outputs from sensors, may give rise to advanced predictive modeling by virtue of synergistic information. Nevertheless, there is a variety of technical and operational challenges in the fusion of multimodal data utilizing CNNs and ML algorithms. These difficulties are mostly due to disparities in data representation, alignments, computational needs, and privacy issues, especially in high-stakes industries such as health care and industrial monitoring. While not addressing these fundamental issues, multimodal fusion systems may succumb to inefficiencies, inaccurate prediction, and limited scalability. This section takes a look at the significant barriers that are impeding the effective fusion of multimodal data.

1. Heterogeneity of Data Formats

The diversity of different types of data is one of the most common issues in multimodal fusion. CNNs are very good at extracting hierarchical features from image data, while structured sensor data and unstructured text need very different preprocessing pipelines and means of modeling. Such a variation makes it difficult to bring together features extracted from different modalities before fusion. For example, image data are usually dense and spatially correlated, sensor data are time-sequenced, while text needs semantic embedding. Under-harmonization of these formats may cause incompatibility of features and weak model convergence (Gao et al., 2020). This variety requires customised fusion architectures that can smartly normalise, encode, and align data types, prior to integration.

2. Synchronization and Alignment Issues

While working with realtime/sequential data streams, temporal synchronization between modalities is critical. Sensor readings, text logs and video frames have to be temporally aligned so as to maintain the contextual relevance. Discrepancies between modalities may cause incoherent predictions and a low level of reliability in the model. Terai et al. (2016) pointed out similar problems in the case of the transcriptomic data, where it was necessary to use sophisticated matching algorithms to align RNA interactions across various databases, because of the asynchrony in processes undertaken to collect data. In multimodal machine learning, approaches like attention mechanisms and dynamic time warping have been suggested to help deal with this problem; however, these solutions usually carry some computational requirements and a higher model complexity.

3. Computational Resource Constraints

Multimodal fusion models are also resource-hungry in that there is a necessity to have parallel data pipelines and joint learning frameworks. Combining CNN-based extraction of features from media of high resolution with real-time ML inferences made on sensor streams adds a large number of bytes to the memory footprint and takes more time to train. According to Chen et al. (2017), even the computationally pure CNN-only systems like Eyeriss need special, energy-efficient hardware to be able to operate in real time. Combining with other modalities, the requirement for computational resources magnifies, and it becomes a challenge to implement such systems on low-power or edge computing-based environments. This presents a real-world practical challenge to large-scale implementation in resource-limited domains such as mobile healthcare and IoT-based agriculture.

4. Overfitting Due to High Dimensionality

Multimodal fusion may only increase the issue of high dimensionality since the concatenating of the features from various modalities tends to provide a huge input space. This leads to the problem of overfitting, especially when the training datasets are few or imbalanced. High-dimensional spaces of the features of Dass et al. (2020) for the purpose of protein disorder prediction, apart from this additional regularization and dimensionality reduction techniques, were also necessary for good generalization. When accuracy in CNN-ML fusion models is high, improper regularization and poor feature selection can lead to the memorization of patterns in the training sets and not generalization in unseen instances. Dropout, L1/L2 regularization, and principal component analysis (PCA) are often used, but parameter selection is not an easy task.

5. Security and Privacy Issues in Healthcare Applications

The use of multimodal fusion in such privacy-sensitive domains as healthcare brings additional risks of privacy and data security. The integration of patient imaging data, physiological sensor inputs, and textual medical records increases the number of possible data leaks and cyber threats by order of magnitude. According to Kaushik et al. (2018), the need to secure medical imaging data was highlighted with the use of such methods as pixel-level protection and genetic algorithms. However, when there are cases where multiple modalities are fused, the traditional security mechanisms might be less effective in offering complete protection. In addition, regulatory frameworks like HIPAA and GDPR have strict rules concerning the storage, transmission, and processing of multimodal health data; this provides for an additional layer of compliance complexity to the deployment of models.

Summing up, although multimodal data fusion approaches based on CNN and ML algorithms provide strong predictive capabilities, these benefits are brought with significant implementation barriers. The control of heterogeneity, synchronization, resource consumption, and overfitting prevention as well as protection of the personal data are important for building effective and ethical multimodal prediction systems.

5. FRAMEWORK FOR MULTIMODAL FUSION OF CNN-ML TECHNOLOGIESATICALLY

Whereas, increasing the complexity of real-world data inevitably results in suboptimal predictions in cases of isolated modality processing, like exploiting only images or text, because a part of the information inevitably gets lost due to the lack of representation by one or the other part of the diverse data. To overcome this, we propose a state-of-the-art multimodal fusion framework that combines the power of Convolutional Neural Networks (CNNs) for image classification with traditional machine learning (ML) techniques for structured sensor (or textual) data. This fusion approach guarantees that every modality provides its unique characteristics in the direction of a unified prediction goal. CNNs are great at hierarchical feature abstraction from image pixels, but when it comes to machine learning techniques like Random Forests, Support Vector Machines (SVM), or Gradient Boosting, they are good at capturing non-linear relationships and feature interactions in structured or temporal data streams. (Jain & Srihari, 2023; Gao et al., 2020). The proposed architecture also caters to the data preprocessing, fusion strategies, and training paradigms that support synchronized learning of modalities. Successful cases of the application of such frameworks have been noticed in multimodal medical diagnostics, quality control of agriculture, and biometric verification systems (Alshehri & AlSaeed, 2022; Mohd Ali et al, 2023; Haghighat et al, 2016).

5.1. Architecture Overview

The architecture has three major components, however. a CNN encoder, a menagerie of ML approaches, and a fusion module:

A. CNN Encoder for Image Data:

Off-the-shelf or tailor-made CNN models (VGG, ResNet, or Eyeriss (Chen et al., 2017)) are used to pull spatial and semantic features out of image-based inputs. These models define hierarchical structures and eliminate

spatial redundancy, therefore, are suitable for studying complex visual information, such as medical scans or thermal images (Alshehri & AlSaeed, 2022; Vaz & Balaji, 2021).

B. ML Algorithms for Sensor/Text Data:

Features that are extracted from sensors or text (patient history, temperature data, or levels of humidity) are passed into any ML model, such as Random Forests, Gradient Boosting, or SVMs. Such models are selected in terms of interpretability, robustness, and ability to work with various data types (Jain & Srihari, 2023). Kono et al., 2011). In the previous studies, the same type of ML algorithms were successfully utilised for the tasks of property prediction and classification for the structured non-visual data sets (Azzam, 2023).

C. Fusion Layer:

The two main fusion strategies are applied:

- Feature-Level Concatenation: Joins vectorized outputs of the CNN and ML models into one feature vector for the ultimate classification (Haghighat et al., 2016).
- Attention-Based Fusion: Uses the mechanisms that dynamically place importance weights on each modality, making context relevance and interpretability better (Lotter et al., 2023).

5.2. Preprocessing Steps

So that integration across modalities is consistent and meaningful, certain preprocessing strategies are applied:

- Image Preprocessing: This includes normalization, resizing, and adjustment of the contrast. Domain specific preprocessing such as thermal calibration can be used in some cases (Mohd Ali et al., 2023).
- Text/Sensor Data Processing: Includes scaling of features (example, min-max or z-score normalization), imputation of missing data, reduction in the number of components using PCA or embedding models (such as TF-IDF or Word2Vec) for different modalities (Xiong et al., 2021; Terai et al., 2016).

5.3. Training Strategy

A mix of training processes is applied; individual modality-specific encoders are pretrained separately and fine-tuned in a joint manner under a joint loss function.

- Modality-Specific Losses: Individual loss terms are used for CNN and ML branches, so that each modality can be optimized efficiently.
- Multi-Loss Optimization: Such losses are compensated through weighting mechanisms during backpropagation so as not one modality can dictate the learning process (Skinnider et al., 2020).
- Early Stopping and Cross-Validation: Strong against overfitting in small and imbalanced data such as those of thermographic scans and patient-reported data (Alshehri & AlSaeed, 2022; Mohd Ali et al., 2023).

5.4. Dataset Examples and Application Scenarios

To validate the proposed framework, there are two representative multimodal datasets that are used:

- Medical Diagnosis Dataset: Integrates breast cancer images created with thermography with patient data and demographic information (Alshehri & AlSaeed, 2022).
- Agricultural Quality Assessment: Combines thermal imaging of pineapples, measures of chemical sensors, and environmental logs for evaluating the quality of fruit (Mohd Ali et al., 2023).

Such examples demonstrate the possibilities of CNN-ML fusion in different areas, allowing for higher prediction accuracy than the unimodal methods.

Component	Model Type	Functionality	Reference
CNN Encoder	Deep CNN	Image feature extraction	Chen et al. (2017); Vaz & Balaji (2021)
ML Predictor	Random Forest / SVM	Structured/sensor data modeling	Jain & Srihari (2023); Azzam (2023)
Fusion Layer	Feature/Aggregation	Combining modality-specific features	Haghighat et al. (2016); Lotter et al. (2023)
Preprocessing	Normalization, Embedding	Input standardization across data types	Terai et al. (2016); Xiong et al. (2021)
Training Module	Multi-loss Strategy	Balanced learning across all data channels	Skinnider et al. (2020)
Application Dataset	Medical, Agriculture	Demonstrating domain versatility	Alshehri & AlSaeed (2022); Mohd Ali et al. (2023)

Table 2: Proposed Framework Component Description

6. CONCLUSION AND FUTURE WORK

It has been observed in recent years that the use of Convolutional Neural Networks (CNNs) with the conventional Machine Learning (ML) algorithms has shown amazing promises in improving multimodal data analysis. Amidst the rising number of heterogeneous data sources, including images, sensor streams, and textual content, intelligent systems that will enable the integration of the modalities to provide holistic and context-aware prediction need to be developed. Although CNNs are good at recognising patterns with spatial properties in the visual data, ML techniques know how to deal with structured and sequential data such as sensor readings or natural language. The coupling of these paradigms brings a strong architecture that can be used in domains like healthcare diagnostics, agricultural monitoring, and environmental sensing (Mohd Ali et al., 2023). Gao et al, 2020, Lotter et al, 2023).

Our work highlights the fact that CNN-ML fusion has a clear effect on enhancing prediction accuracy as a result of the exploitation of complementary features from various data modalities. For instance, in thermal imaging for fruit quality classification, the combination of CNN-derived visual attributes with the temperature and moisture sensor data has increased the reliability of prediction (Mohd Ali et al., 2023). Equally, in the medical settings, the combination of CNN-based visual diagnostics with clinical potings or the biosensor results has been established to increase the rates of detection of the diseases (Alshehri & AlSaeed, 2022; Vaz & Balaji, 2021). This synergy of deep learning and classical ML methods helps avoid reliance on specific data types, hence enhancing generalization of the models whilst being resistant to missing or corrupted modality inputs (Haghighat et al., 2016; Yan et al., 2015).

Furthermore, this approach to fusion allows real-time analytics where CNN evaluates visual feeds in a lightning-quick manner, while ML algorithms reason in real-time with structured sensor outputs or historical datasets to tune predictions on-the-fly (Chen et al., 2017; Wunsch et al., 2021). The opportunity to combine several knowledge streams renders these hybrid systems quite appropriate for the critical nature of their applications, including smart health monitoring and autonomous environmental sensing (Lotter et al., 2023 p; Terai et al., 2016).

6.1. Future Research Directions

There are some promising extensions that can further develop this field in the future:

- Incorporating attention-based transformers: The adoption of the transformer architectures (e.g., Vision Transformers (ViT) and cross-modal attention mechanisms) in the fusion pipelines could help with a better context alignment between modalities. Such an approach will make the model selective and focus on the important data streams of reducing interpretation and efficiency, especially in high-dimensional input-based applications such as genomics or neuroimaging (Skinnider et al., 2020; Scholkopf et al., 2021).
- Real time multimodal systems for smart healthcare and agriculture: The future work will focus on the development of lightweight deployable multimodal system able to process real time data streams. In medical care, this would allow for constant monitoring of patient with a simultaneous fusion of camera flows and biosensor information. In agriculture, a combination of drone imagery, soil sensors, and climatic readings may maximize the estimates of crop well-being and control of irrigation (Mohd Ali et al., 2023; Xiong et al., 2021).
- Building up fusion to multi-sensor + text + image frameworks: Most of the current research is devoted to two-modal fusion (image+text, image+sensor). The next frontier would be creating strong models that could work in concert with three or more modalities, concurrently – for example, integrating patient scan images, sensor readings, and electronic medical records for richer diagnostics (Jain & Srihari, 2023; Zhou et al., 2016). This would require developing higher-level mechanisms for aligning things, perhaps graphical or hierarchical fusion models (Gao et al., 2020).

Finally, a combination of CNNs and ML techniques opens the way towards smart, scalable, and comprehensive prediction engines in a broad range of real-world settings. Uncovering the shortcomings of the current usage of the single modality paradigm and turning to attention-based, real-time, and multi-source integration, the field is ready for major innovation and impact.

REFERENCES

[1] Alshehri, A., & AlSaeed, D. (2022). Breast Cancer Detection in Thermography Using Convolutional Neural Networks (CNNs) with Deep Attention Mechanisms. Applied Sciences (Switzerland), 12(24). https://doi.org/10.3390/app122412922

[2] Azzam, K. A. (2023). SwissADME and pkCSM Webservers Predictors: an integrated Online Platform for Accurate and

Comprehensive Predictions for In Silico ADME/T Properties of Artemisinin and its Derivatives. Kompleksnoe Ispol′zovanie Mineral′nogo Syr′â/Complex Use of Mineral Resources/Mineraldik Shikisattardy Keshendi Paidalanu, 325(2), 14–21. https://doi.org/10.31643/2023/6445.13

[3] Chen, Y. H., Krishna, T., Emer, J. S., & Sze, V. (2017). Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks. IEEE Journal of Solid-State Circuits, 52(1), 127–138. https://doi.org/10.1109/JSSC.2016.2616357

[4] Terai, G., Iwakiri, J., Kameda, T., Hamada, M., & Asai, K. (2016). Comprehensive prediction of lncRNA-RNA interactions in human transcriptome. BMC Genomics, 17(1). https://doi.org/10.1186/s12864-015-2307-5

[5] Dass, R., Mulder, F. A. A., & Nielsen, J. T. (2020). ODiNPred: comprehensive prediction of protein order and disorder. Scientific Reports, 10(1). https://doi.org/10.1038/s41598-020-71716-1

[6] Haghighat, M., Abdel-Mottaleb, M., & Alhalabi, W. (2016). Discriminant Correlation Analysis: Real-Time Feature Level Fusion for Multimodal Biometric Recognition. IEEE Transactions on Information Forensics and Security, 11(9), 1984–1996. https://doi.org/10.1109/TIFS.2016.2569061

[7] Kaushik, P., & Jain, M. (2018). Design of low power CMOS low pass filter for biomedical application. International Journal of Electrical Engineering & Technology (IJEET), 9(5). https://iaeme.com/MasterAdmin/Journal_uploads/IJEET/VOLUME_9_ISSUE_5/IJEET_09_05_003.pdf

[8] Kono, N., Arakawa, K., & Tomita, M. (2011). Comprehensive prediction of chromosome dimer resolution sites in bacterial genomes. BMC Genomics, 12. https://doi.org/10.1186/1471-2164-12-19

[9] Kaushik, P., Jain, M., & Jain, A. (2018). A pixel-based digital medical images protection using genetic algorithm. Int J Electron Commun Eng, 11, 31-37. http://www.irphouse.com/ijece18/ijecev11n1_05.pdf

[10] Skinnider, M. A., Johnston, C. W., Gunabalasingam, M., Merwin, N. J., Kieliszek, A. M., MacLellan, R. J., … Magarvey, N. A. (2020). Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. Nature Communications, 11(1). https://doi.org/10.1038/s41467-020-19986-1

[11] Kaushik, P., & Jain, M. A Low Power SRAM Cell for High Speed Applications Using 90nm Technology. Csjournals. Com, 10. https://www.researchgate.net/publication/391458245_A_Low_Power_SRAM_Cell_for_High_Speed

[12] Gao, J., Li, P., Chen, Z., & Zhang, J. (2020, May 1). A survey on deep learning for multimodal data fusion. Neural Computation. MIT Press Journals. https://doi.org/10.1162/neco_a_01273

[13] KAUSHIK, P., JAIN, M., & SHAH, A. (2018). A Low Power Low Voltage CMOS Based Operational Transconductance Amplifier for Biomedical Application. https://ijsetr.com/uploads/136245IJSETR17012-283.pdf

[14] Terai, G., Iwakiri, J., Kameda, T., Hamada, M., & Asai, K. (2016). Comprehensive prediction of lncRNA-RNA interactions in human transcriptome. BMC Genomics, 17(1). https://doi.org/10.1186/s12864-015-2307-5

[15] Puneet Kaushik, Mohit Jain , Gayatri Patidar, Paradayil Rhea Eapen, Chandra Prabha Sharma (2018). Smart Floor Cleaning Robot Using Android. International Journal of Electronics Engineering. https://www.csjournals.com/IJEE/PDF10-2/64.%20Puneet.pdf

[16] Mohd Ali, M., Hashim, N., Abd Aziz, S., & Lasekan, O. (2023). Utilisation of Deep Learning with Multimodal Data Fusion for Determination of Pineapple Quality Using Thermal Imaging. Agronomy, 13(2). https://doi.org/10.3390/agronomy13020401

[17] Lotter, L. D., Kohl, S. H., Gerloff, C., Bell, L., Niephaus, A., Kruppa, J. A., … Konrad, K. (2023, March 1). Revealing the neurobiology underlying interpersonal neural synchronization with multimodal data fusion. Neuroscience and Biobehavioral Reviews. Elsevier Ltd. https://doi.org/10.1016/j.neubiorev.2023.105042

[18] Mohit Jain , Adit Shah "Machine Learning with Convolutional Neural Networks (CNNs) in Seismology for Earthquake Prediction" Iconic Research And Engineering Journals Volume 5 Issue 8 2022 Page 389-398

https://www.irejournals.com/index.php/paper-details/1707057

[19] Wunsch, A., Liesch, T., & Broda, S. (2021). Groundwater level forecasting with artificial neural networks: A comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX). Hydrology and Earth System Sciences, 25(3), 1671–1687. https://doi.org/10.5194/hess-25-1671-2021

[20] Vaz, J. M., & Balaji, S. (2021). Convolutional neural networks (CNNs): concepts and applications in pharmacogenomics. Molecular Diversity, 25(3), 1569–1584. https://doi.org/10.1007/s11030-021-10225-3

[21] Jain, M., & None Arjun Srihari. (2023). House price prediction with Convolutional Neural Network (CNN). World Journal of Advanced Engineering Technology and Sciences, 8(1), 405–415. https://doi.org/10.30574/wjaets.2023.8.1.0048 https://wjaets.com/sites/default/files/WJAETS-2023-0048.pdf

[22] Toufigh, V., & Jafari, A. (2021). Developing a comprehensive prediction model for compressive strength of fly ash-based geopolymer concrete (FAGC). Construction and Building Materials, 277. https://doi.org/10.1016/j.conbuildmat.2021.122241

[23] Xiong, G., Wu, Z., Yi, J., Fu, L., Yang, Z., Hsieh, C., … Cao, D. (2021). ADMETlab 2.0: An integrated online platform for accurate and comprehensive predictions of ADMET properties. Nucleic Acids Research, 49(W1), W5–W14. https://doi.org/10.1093/nar/gkab255

[24] Yan, X., Kang, W., Deng, F., & Wu, Q. (2015). Palm vein recognition based on multi-sampling and feature-level fusion. Neurocomputing, 151(P2), 798–807. https://doi.org/10.1016/j.neucom.2014.10.019

[25] Zhou, T., Lu, H., Zhang, J., & Shi, H. (2016). Pulmonary nodule detection model based on SVM and CT image feature-level fusion with rough sets. BioMed Research International, 2016. https://doi.org/10.1155/2016/8052436