Capstone Project Report

Deep Learning - Facial Emotion Detection

Mohit Pammu

Executive Summary

This project introduces **CNN (Convolutional Neural Network) Model 3**, a deep learning model designed to accurately detect and classify human emotions based on facial expressions. Model 3 leverages advanced data augmentation techniques, multiple convolutional layers, and dropout for regularization, enabling a strong balance between model complexity, computational efficiency, and generalization to unseen data. The dataset used is grayscale, and experiments confirmed that models trained on grayscale images consistently outperformed those using RGB inputs. This not only aligns with the nature of the data but also enhances processing efficiency.

Model 3 delivered the highest performance, achieving approximately **~82% test accuracy**—outperforming both simpler CNN architectures and transfer learning models. Despite its success, Model 3 faced several challenges:

- **Ambiguity in labeling**: Emotions like ‘sad’ and ‘neutral’ share overlapping features, leading to occasional misclassifications.
- **Class imbalance**: Reliance on licensed datasets introduced uneven representation across emotion classes.
- **Technological limitations**: Static grayscale images, while efficient, limit the model's application in dynamic, real-time scenarios.

These insights pave the way for future enhancements, including richer datasets, improved label clarity, and integration with video-based real-time systems.

Problem Summary

As artificial intelligence (AI) continues to evolve, enabling machines to interpret and respond to human emotions—known as **Affective Computing**—has become a pivotal frontier. Facial expression recognition (FER) is central to this endeavor, given that over **55% of human emotional communication is conveyed through facial cues** (Mehrabian, 1971; WSJ, 2024).

The global **Affective Computing market**, valued at **$62.53 billion in 2023**, is projected to expand at a **CAGR of 30.6%**, **reaching $388.28 billion by 2030** (Grand View Research, 2024). This growth is driven by the integration of emotion-aware technologies across sectors such as healthcare, automotive, education, and customer service. For instance, in healthcare, FER aids in early detection of mental health conditions, while in automotive, it enhances driver safety by monitoring fatigue and stress levels.

Despite these advancements, challenges persist. Recent studies highlight that **facial expressions are not universally indicative of specific emotions**, as cultural and contextual factors significantly influence emotional expression (Barrett et al., 2019; WSJ, 2024). Moreover, concerns about **privacy and ethical implications** arise, especially with the deployment of FER technologies in surveillance and public spaces (Axios, 2019).

To address these challenges, this project aims to develop a **deep learning-based computer vision model** capable of accurately classifying facial expressions into distinct emotional categories. By leveraging advanced data augmentation techniques and robust neural network architectures, the model seeks to achieve high accuracy while ensuring generalizability across diverse populations and contexts.

Successfully implementing such a model holds the potential to revolutionize human-computer interactions, making them more empathetic and responsive. It can enhance user experiences in virtual assistants, improve patient care in telemedicine, and contribute to safer autonomous driving systems. However, it is imperative to balance technological capabilities with ethical considerations to ensure responsible and equitable deployment of FER technologies.

Solution Design

After evaluating multiple models—including two custom grayscale CNNs and three RGB-based transfer learning architectures (VGG16, ResNet101, EfficientNetV2B2)—Model 3, a deeper grayscale CNN, was selected for deployment. It achieved the highest test accuracy (~82%) displayed in **Figure 1** and demonstrated strong generalization on unseen data while maintaining reasonable computational efficiency. Model 3's specifications include:

- **Three Convolutional Blocks** with increasing filter sizes (32, 64, 128) for hierarchical feature extraction
- **Dual Convolutions per Block** for richer representations before down sampling
- **Batch Normalization** after each convolution to stabilize and accelerate training
- **MaxPooling Layers** to reduce spatial dimensions
- **Strategic Dropout:** 25% after conv layers, 50% after dense layers to mitigate overfitting
- **Dense Layers:** 256 and 128 neurons for high-level abstraction
- **Softmax Output:** 4 neurons for multi-class classification

Training Setup:

- **Loss Function**: Categorical Cross-Entropy (aligned with softmax)
- **Optimizer**: Adam (LR = 0.001)
- **Batch Size**: 32
- **Epochs**: 50 with early stopping (patience = 15)
- **Data Augmentation**: rotation, shifts, zoom, brightness, and flipping for robustness

Model 3's balance of performance and efficiency makes it well-suited for **real-time, edge-friendly facial emotion detection** applications. Compared to transfer learning models trained on RGB datasets (e.g., ImageNet), this grayscale-optimized CNN avoids unnecessary parameter inflation and noise introduced by irrelevant channels.

Industry Relevance:

- Lightweight CNNs are increasingly preferred for edge deployment where compute resources are limited (Jain et al., 2022).
- Custom grayscale models are effective in domains like **education tech**, **telemedicine**, and **customer sentiment analysis** where fast, low-power inference is crucial.

- Recent studies (e.g., Mollahosseini et al., 2016) show emotion recognition models that align closely with data modality (e.g., grayscale) perform better than generic transfer-learned architectures.

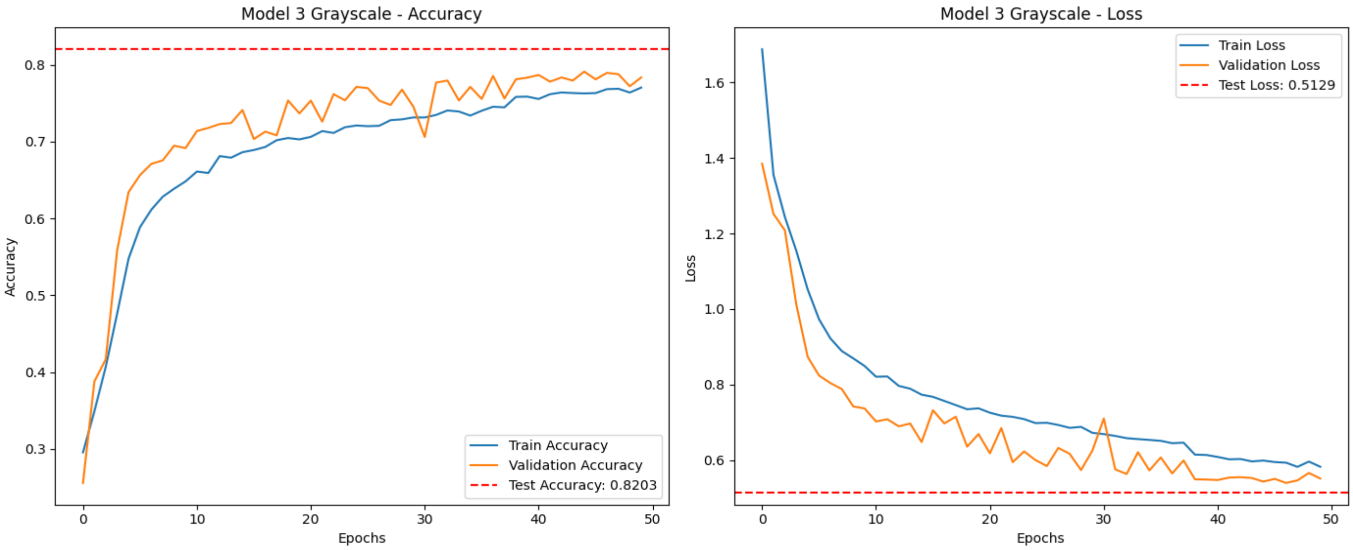


Figure 1: complex CNN Model 3 Grayscale

Below is **Figure 2**, a detailed chart comparing model performances.

Model	Architecture Type	Input Mode	Test Accuracy	Training Time per Epoch	Notes
Model 1 (GS)	Simple CNN	Grayscale	66%	~11 sec	Basic architecture; limited feature extraction
Model 1 (RGB)	Simple CNN	RGB	66%	~11 sec	RGB input added no performance benefit
Model 2 (GS)	Medium CNN	Grayscale	72%	~64 sec	Better accuracy, slight overfitting observed
Model 2 (RGB)	Medium CNN	RGB	66%	~64 sec	RGB version underperformed vs grayscale
Model 3	**Complex CNN**	**Grayscale**	**82%**	**~44 sec**	**Best performance, good generalization, efficient and effective**
VGG16	Transfer Learning	RGB	51%	~67 sec	Poor performance on grayscale task; large model
ResNet101	Transfer Learning	RGB	25%	High	Computationally expensive; underfitting on grayscale
EfficientNetV2B2	Transfer Learning	RGB	25%	~20 sec	Fast but highly inaccurate; poor feature alignment with task

Figure 2: Model comparison overview

Analysis and Key Insights:

This study reveals critical takeaways for industries adopting emotion recognition systems. Standard transfer learning models such as VGG16 and ResNet underperformed due to their reliance on RGB features, which mismatched the grayscale nature of the dataset. In contrast, custom CNN architectures tailored for grayscale input achieved significantly higher accuracy and computational efficiency—highlighting a clear advantage for deploying lightweight, task-specific models in resource-constrained or real-time environments (Zhao et al., 2021).

Data augmentation proved effective in addressing class imbalance, especially improving detection of underrepresented emotions like surprise. This underscores the broader applicability of augmentation strategies in domains with uneven data distribution—ranging from behavioral analytics and fraud detection to diagnostic imaging—where synthetic data can enhance generalization and reduce overfitting without costly new data acquisition (Shorten & Khoshgoftaar, 2019).

Persistent confusion between visually similar classes such as sad and neutral exposed a key limitation of static image-based models: their inability to capture fine-grained temporal cues. This suggests that future systems—particularly those used in emotionally sensitive applications like teletherapy, autism support, or interactive education—would benefit from integrating video-based analysis or multi-modal signals (e.g., voice tone, posture) to enhance emotional nuance and system reliability.

Finally, the results emphasize the importance of **designing with data characteristics in mind**. General-purpose deep learning models may not translate effectively across domains without adaptation. Instead, domain-aligned architectures—like the grayscale-optimized CNN used here—offer more effective, scalable, and interpretable solutions across sectors that rely on emotion-aware systems.

Challenges and Limitations

A critical challenge in developing the model was **class imbalance**, especially the underrepresentation of the *surprise* emotion class. This led to biased predictions and reduced classification accuracy for minority classes. Although data augmentation provided some improvement, it fell short of simulating real-world variability. This limitation is especially relevant in industry domains such as **mental health monitoring** or **sentiment analysis**, where consistent detection across all emotional categories is crucial.

Another persistent issue was the **confusion between visually similar emotions**, such as *sad* and *neutral*, as reflected in the confusion matrix (Appendix 3). Static grayscale images limited the model's ability to distinguish subtle facial expressions. This points to the need for **temporal modeling** (e.g., video sequences) or **multi-modal inputs** (e.g., combining audio, physiological signals, or text) in emotionally nuanced applications like **customer support** and **human-computer interaction**.

Additionally, **pre-trained transfer learning models** demonstrated poor performance in this context due to their reliance on RGB image inputs and general-purpose feature representations. This reinforces the need for **domain-specific model design**, especially for grayscale datasets or context-specific emotion detection. Future solutions should consider fine-tuning pre-trained networks with relevant grayscale datasets, integrating attention mechanisms, or applying spatiotemporal architectures to better capture emotional nuance.

Beyond technical constraints, deploying emotion recognition systems in the real world demands **strict compliance with global and sector-specific data protection regulations**. For example:

- Under **GDPR (EU)** and **CCPA/CPRA (California)**, processing biometric data like facial expressions requires a defined legal basis, clear consumer rights mechanisms, privacy notices, and comprehensive documentation (including DPIAs and data inventories).
- **Sector-specific compliance** is equally important:
    - In **healthcare**, adherence to **HIPAA** is required for systems that handle patient emotional data, including clinical validation if used diagnostically.
    - In **education**, systems must comply with **FERPA**, include safeguards for child data, and align with research ethics and age-appropriate standards.
    - In **financial services**, compliance involves explainability standards, anti-discrimination safeguards, and auditability—especially where emotion AI might influence decision-making.

To operationalize this, organizations must implement **compliance frameworks** with ongoing documentation, regular audits, regulatory change monitoring, and training for deployment teams. **Incident response protocols** are also essential to mitigate risks in the event of compliance breaches.

In summary, while the proposed solution demonstrates strong technical performance, especially through Model 3's domain-aligned design, its real-world viability hinges on addressing dataset limitations, enhancing modality richness, and embedding robust privacy and regulatory safeguards from the outset.

Recommendations for Implementation

To operationalize the proposed solution, stakeholders should adopt **Model 3** as the production model. This custom CNN architecture is optimized for grayscale input and achieved the highest test accuracy (~82%) with strong generalization and minimal overfitting. Its efficient training time (~44 seconds per epoch) makes it computationally cost-effective for both development and scaling. To further enhance performance, especially for harder-to-distinguish emotions like 'sad' and 'neutral', techniques such as **targeted data augmentation**, **weighted loss functions** (which assign more importance to underrepresented classes), and **attention mechanisms** (which help the model focus on key facial regions) should be explored.

Investing in **scalable infrastructure** is critical for supporting both training and deployment. Cloud-based platforms (e.g., AWS, GCP) are recommended for their flexibility and cost-effectiveness, while **edge deployment** options should be considered for real-time applications like kiosks or mobile devices. Additionally, stakeholders should allocate resources for expanding the dataset, particularly by sourcing more diverse, real-world facial data to reduce demographic and situational bias. This can be achieved through partnerships, open-source datasets, or controlled data collection initiatives, with appropriate consent and privacy safeguards.

To ensure long-term accuracy and ethical use, a **continuous improvement pipeline** should be established. This includes monitoring model performance in production, retraining with new data, and incorporating feedback from end users. Collaborating with **domain experts**—such as psychologists or behavioral scientists—can refine the emotional labeling process and improve prediction reliability. Ethical considerations are essential, especially in sensitive areas like mental health or security, and should be addressed through **bias audits**, transparency in model decisions, and user consent protocols.

Key risks include persistent class imbalance, misclassification of visually similar emotions, and scalability under high data loads. These can be mitigated through **real-world testing**, **model auditing**, and exploring **multi-modal enhancements** (e.g., integrating voice tone or video sequences) to capture a fuller emotional context. These steps will ensure that the solution is not only accurate but also robust, ethical, and ready for deployment in real-world settings.

Bibliography

- Allen-Ebrahimian, B. (2019, November 2). *China's emotion-recognition tech raises alarm*. Axios. https://www.axios.com/2019/11/02/china-emotion-recognition

- Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest, 20*(1), 1–68. https://doi.org/10.1177/1529100619832930

- Grand View Research. (2024). *Affective computing market size, share & trends analysis report by technology (touch-based, touchless), by end use, by region, and segment forecasts, 2024 - 2030*. https://www.grandviewresearch.com/industry-analysis/affective-computing-market

- Jain, R., Raj, R., & Patel, H. (2022). Lightweight deep learning architectures for edge computing: Trends and challenges. *IEEE Access, 10*, 110236–110251. https://doi.org/10.1109/ACCESS.2022.3211845

- Mehrabian, A. (1971). *Silent messages*. Wadsworth Publishing.

- Mims, C. (2024, February 6). Think AI can perceive emotion? Think again. *The Wall Street Journal*. https://www.wsj.com/tech/ai/think-ai-can-perceive-emotion-think-again-2b4c7d29

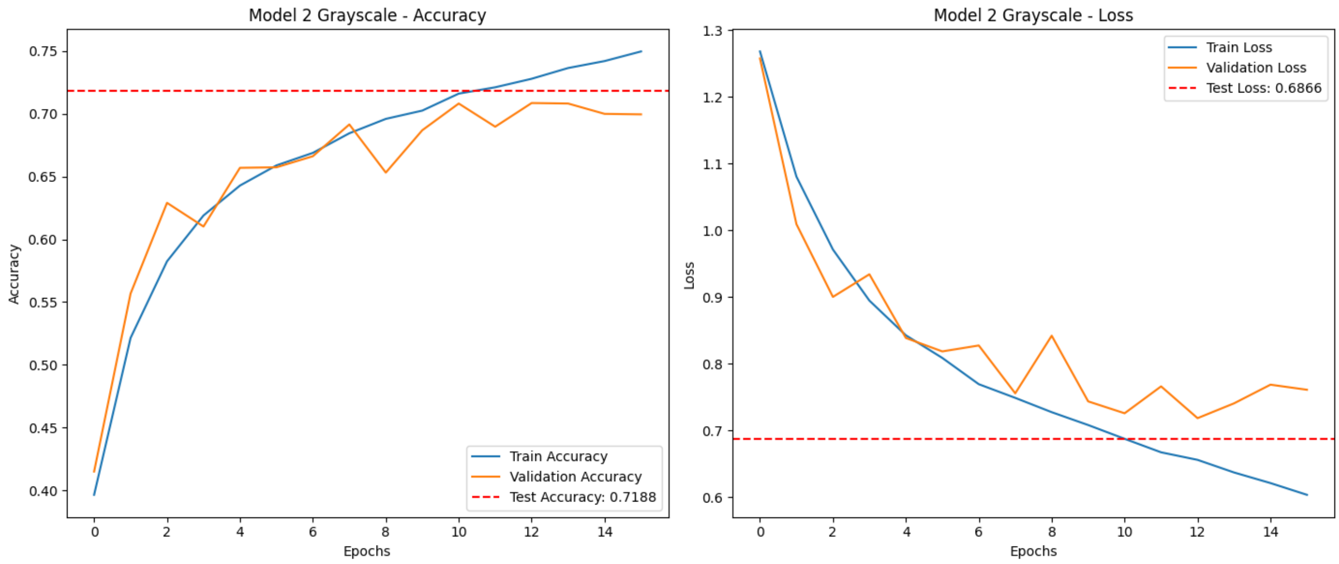
- Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2016). AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing, 10*(1), 18–31. https://doi.org/10.1109/TAFFC.2017.2740923

- Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), 60. https://doi.org/10.1186/s40537-019-0197-0

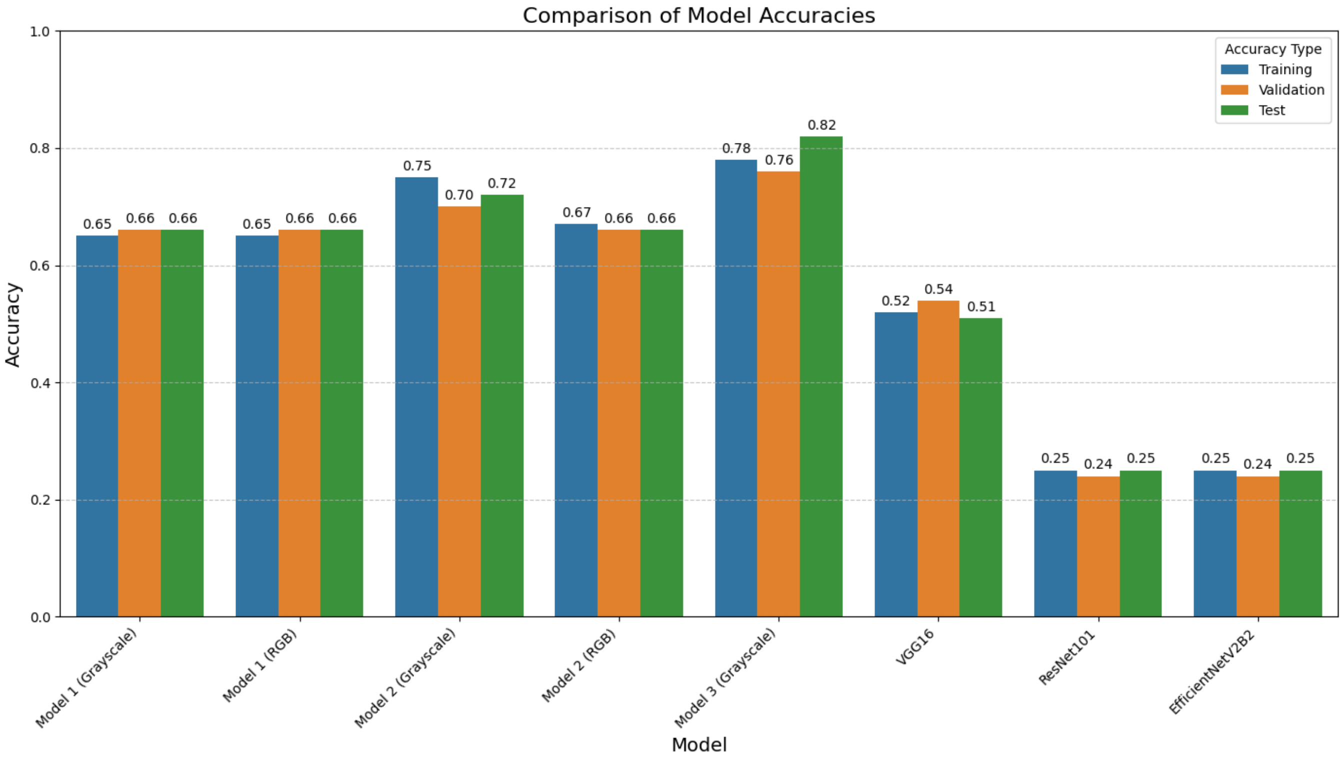
- Zhao, X., Li, S. Z., & Zhang, Y. (2021). Efficient facial recognition on grayscale images: A comparison of neural architectures. *Pattern Recognition Letters*, 143, 31–37. https://doi.org/10.1016/j.patrec.2020.12.008

Appendix

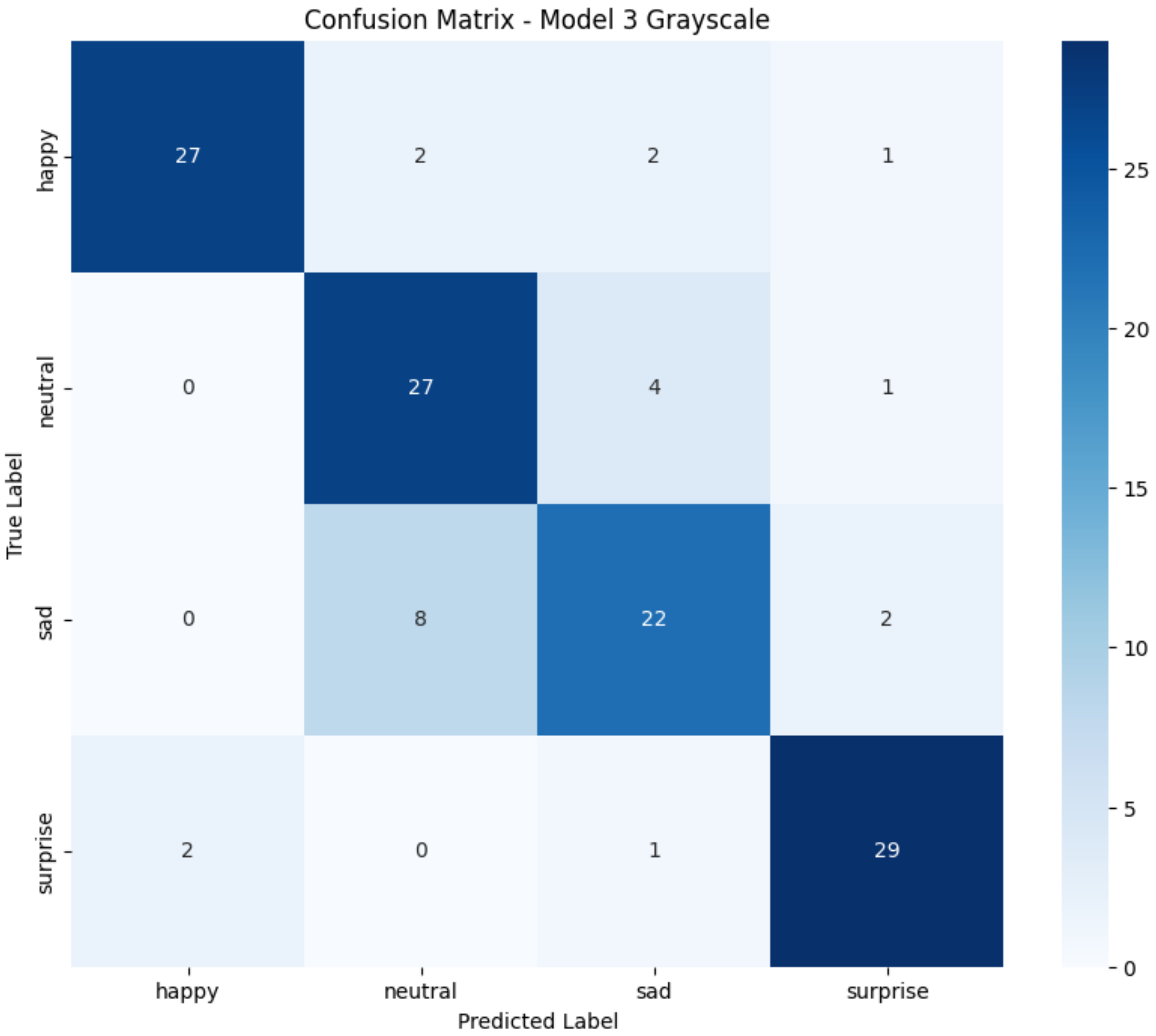
Appendix 1: CNN Model 2 Grayscale



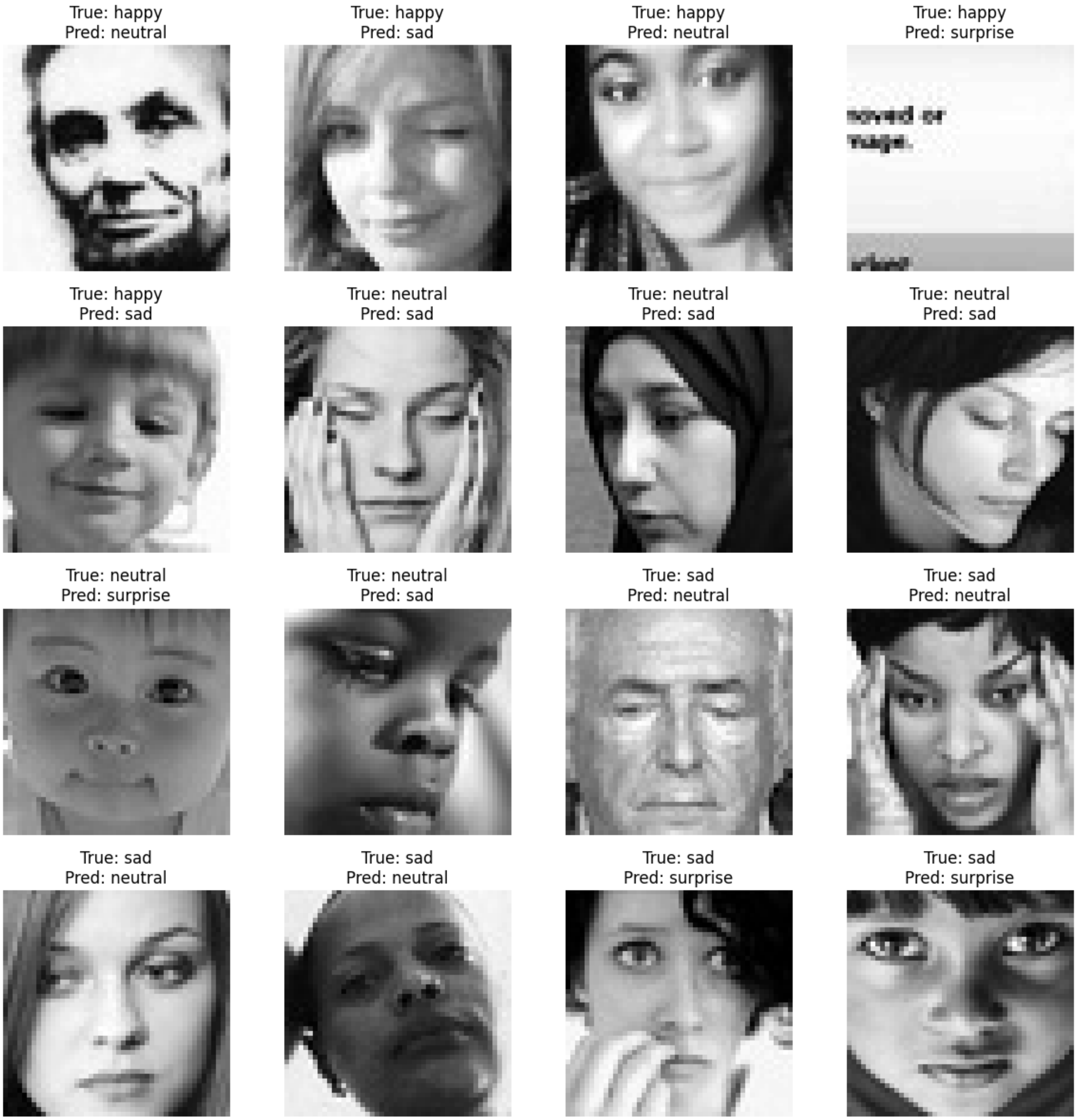
Appendix 2: Model comparison graph



Appendix 3: Confusion matrix for Model 3



Appendix 4: Model 3 Misclassifications



True: happy
Pred: neutral

True: happy
Pred: sad

True: happy
Pred: neutral

True: happy
Pred: surprise

True: happy
Pred: sad

True: neutral
Pred: sad

True: neutral
Pred: sad

True: neutral
Pred: sad

True: neutral
Pred: surprise

True: neutral
Pred: sad

True: sad
Pred: neutral

True: sad
Pred: neutral

True: sad
Pred: neutral

True: sad
Pred: neutral

True: sad
Pred: surprise

True: sad
Pred: surprise