It is imperative to adopt a multifaceted assessment strategy when evaluating the performance of convolutional-recurrent neural network architectures for OCR tasks, particularly given the complexities inherent in early modern printed texts. Our objective transcends mere accuracy; we aim to holistically understand model behaviour across various performance dimensions, acknowledging that the idiosyncrasies of historical texts pose unique challenges.

1. Character and Word Recognition Accuracy:

At the foundational level, accuracy remains a pivotal metric, quantifying the proportion of characters or words correctly recognised by the model. However, given the nuanced nature of our task—where ancient spellings, typographical variations, and OCR-induced artefacts interplay—mere accuracy might not fully encapsulate model efficacy. Therefore, while I propose to utilise accuracy as a baseline metric, it will be contextualised within a broader evaluative framework.

2. Precision, Recall, and F1 Score:

To delve deeper into the model’s performance, especially to understand its behaviour in the context of false positives and false negatives—which are critical in OCR tasks—I will employ precision (the proportion of true positive results in all positive predictions) and recall (the proportion of true positive results in all actual positives). The F1 score, which harmonically averages precision and recall, will serve as a unified metric to balance the trade-off between them, providing a singular lens through which to assess the model's robustness, especially in distinguishing between closely similar characters or words.

3. Levenshtein Distance (Edit Distance):

Given the historical nature of the text, with inherent spelling variability and OCR errors, the Levenshtein distance becomes particularly salient. This metric measures the minimum number of edits required to change one string into the other, thus offering a nuanced view of the model's performance in capturing the linguistic essence of the text, beyond mere character or word accuracy. It affords a sensitivity to the types of errors (substitutions, deletions, insertions) that are particularly prevalent in OCR tasks involving historical documents.

4. Confusion Matrix Analysis:

To further elucidate the model's performance on a granular level, I propose conducting a detailed confusion matrix analysis. This will not only highlight the characters or words that are most commonly misrecognised but also expose systemic biases or weaknesses in the model, such as confusing visually similar characters. Such an analysis is instrumental in iterative model refinement, allowing for targeted improvements in character recognition capabilities.

5. Cross-validation and Generalization Performance:

Lastly, to ensure the model’s robustness and ability to generalise across unseen data, I will utilise k-fold cross-validation. This involves partitioning the data into k subsets, training the model on k-1 of these subsets, and validating it on the remaining subset. This process is repeated k times, with each subset serving as the validation set once. Analysing performance across these folds provides insights into the model's consistency and its susceptibility to overfitting, underscoring its generalisation capability to new, unseen data.

In conclusion, by embracing a comprehensive and multidimensional approach to evaluating model performance, we aim to advance the state of OCR technology for historical texts and contribute nuanced insights into the complexities of language evolution and typographical variation over time.