The paper "Deep Neural Networks and Tabular Data: A Survey" by Borisov et al. (2022) provides a comprehensive comparison between traditional machine learning (ML) models and deep learning approaches, specifically artificial neural networks (ANNs), for tabular data. Let's focus on the approach taken by the authors in Sections VII (Experiments) and VIII (Discussion and Future Prospects) and compare it with my approach in the Tabular Data Preprocessing assignment.

Approach in the Paper:

1. Dataset Selection:
   * The authors carefully selected five real-world tabular datasets with diverse characteristics, such as data domain, target variable type (classification or regression), number of categorical and continuous variables, and sample size.
   * The selected datasets covered a range of sizes from small to large, ensuring a comprehensive evaluation.
2. Data Preprocessing:
   * The authors applied consistent preprocessing steps to all datasets and models.
   * They used zero-mean, unit-variance normalization for numerical features and ordinal encoding for categorical features.
   * Missing values were substituted with zeros for linear regression and models based on pure neural networks.
3. Model Selection and Hyperparameter Tuning:
   * The authors included a wide range of ML models and ANNs in their comparison.
   * They used the Optuna library with 100 iterations for each model to tune hyperparameters.
   * Each hyperparameter configuration was cross-validated with five folds.
4. Evaluation Metrics:
   * The authors used accuracy and area under the ROC curve (AUC) as evaluation metrics for classification tasks.
   * For regression tasks, they used mean squared error (MSE).
5. Results and Discussion:
   * The authors found that traditional ML models, particularly gradient boosting decision trees (XGBoost, CatBoost), outperformed deep learning approaches on most datasets, except for the very large HIGGS dataset.
   * They discussed the challenges of deep learning on tabular data, such as low-quality training data, missing or complex irregular spatial dependencies, dependency on preprocessing, and the importance of single features.

Comparison with My Approach:

1. Dataset Selection:
   * In my assignment, I worked with a single dataset related to patient records and metastatic cancer diagnosis.
   * While the paper covered a diverse range of datasets, my focus was on a specific domain and problem.
2. Data Preprocessing:
   * Similar to the paper, I performed data preprocessing steps such as handling missing values and scaling the features using StandardScaler.
   * I removed columns with more than 30% missing values and replaced the remaining missing values with the mean for numerical columns and used one-hot encoding for categorical columns.
3. Model Selection and Evaluation:
   * I trained three different models: Logistic Regression, Random Forest, and Support Vector Machine (SVM).
   * I evaluated the models using the classification report, which provided precision, recall, and F1-score for each class, as well as overall accuracy.
   * The paper used a more extensive set of models and evaluation metrics, including hyperparameter tuning and cross-validation.
4. Dimensionality Reduction:
   * In my assignment, I applied Principal Component Analysis (PCA) for dimensionality reduction.

In summary, while the paper conducted a comprehensive comparison of ML models and ANNs across diverse tabular datasets, my assignment focused on a specific dataset and problem. I applied similar preprocessing techniques and trained multiple models, but the paper's approach was more extensive in terms of dataset diversity, model selection, hyperparameter tuning, and evaluation metrics. The paper's findings highlight the competitiveness of traditional ML models, particularly gradient boosting decision trees, compared to deep learning approaches for tabular data.