External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms

Answers for the Design section questions given in the Rubric:

1. Workflow of the experiment:
   - The workflow involves the selection of mammographic examinations from a public screening program. These examinations are then independently assessed by three different AI CAD algorithms and radiologists. The AI CAD algorithms process the images and provide a prediction score for each breast. The radiologists also make binary decisions regarding the examinations. The outcomes of these assessments are then compared to evaluate the performance of the AI algorithms compared to radiologists.

2. Analysis of outcomes:
   - The outcomes of the experiments are analyzed using various diagnostic metrics such as sensitivity, specificity, accuracy, abnormal interpretation rate, cancer detection rate, false-negative rate, and positive predictive value. These metrics are calculated for both the AI CAD algorithms and radiologists' assessments. Statistical tests, such as the DeLong method, are used to assess differences in performance between the algorithms and radiologists.

3. Patient and doctor selection:
   - Patients: The study sample includes women aged 40 to 74 years who were diagnosed with breast cancer between 2008 and 2015, had a complete screening examination prior to diagnosis, had no prior breast cancer, and did not have implants. Additionally, a random sample of healthy women aged 40 to 74 years was included.
   - Doctors: The examinations are assessed by double-reading by different radiologists. There were 25 different first-reader radiologists and 20 different second-reader radiologists involved in the study. The second reader is often more experienced than the first reader.

4. Proposed benefit of the hybrid system:
   - The proposed benefit of the hybrid system, combining AI CAD algorithms with radiologists' assessments, is to improve the accuracy and efficiency of breast cancer screening. While one AI algorithm performed independently with high diagnostic performance, combining it with radiologists' assessments further increased the number of positive cancer cases detected.

5. Short-term and long-term effects on doctors:
   - Short-term effects: The study does not explicitly mention short-term effects on doctors. However, the introduction of AI CAD algorithms may lead to changes in workflow and potentially require adaptation by radiologists in integrating AI assessments into their practice.
   - Long-term effects: Over the long term, the integration of AI CAD algorithms into breast cancer screening may lead to changes in the role of radiologists, potentially enhancing their efficiency by assisting in the interpretation of mammograms and improving overall screening accuracy. However, it may also necessitate ongoing training and adaptation to new technologies. <This message was edited>

Answer to Evidence section questions given in the rubric:

1. Size of the data used for evaluation (test data):
   - The study used a final study sample consisting of 8805 women and screening examinations. Of these, 739 women received a diagnosis of breast cancer, and 8066 women were healthy controls. This dataset size represents the sample used for evaluation.

2. Diversity of the data collection process:
   - The data collection process includes mammographic examinations from a public mammography screening program, the CSAW (Swedish Cohort of Screen-Age Women) dataset, which covers women aged 40 to 74 years in Stockholm county. The screening interval, age range, and geographical location provide some diversity in the dataset. However, the study does not explicitly mention additional factors such as racial or ethnic diversity, socioeconomic status, or geographic location beyond Stockholm county. Therefore, while the dataset includes a range of ages and screening outcomes, further information is needed to determine the full extent of diversity in the data collection process.

Answers to Claims Questions given in the Rubric:

1. Main Claims:
   - The best-performing AI CAD algorithm (AI-1) demonstrated a diagnostic performance comparable to or surpassing that of radiologists in assessing screening mammograms.
   - AI-1 exhibited an overall AUC of 0.956 for cancer detection at screening or within 12 months thereafter, with a sensitivity of 81.9% and specificity of 96.6%.
   - Combining the first reader with AI-1 identified more cancer cases than combining the first and second readers.
   - AI-1 outperformed the other two algorithms (AI-2 and AI-3) across all subgroups.
   - Combining all three AI algorithms did not significantly improve performance over using the best algorithm alone.

2. Real World vs. Specific Dataset:
   - The claims are about the performance of AI CAD algorithms in assessing screening mammograms, which suggests applicability to the real world. However, the findings are based on a specific dataset and study population, so there could be variations when applied to different populations or settings.

3. Clarity in Title and Abstract:
   - The claims are reasonably clear in the abstract and could be inferred from the title, which mentions "Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms." However, specific performance metrics like AUC, sensitivity, and specificity are not explicitly mentioned in the title or abstract.

4. Effort Required to Reproduce:
   - The effort required to reproduce would depend on the availability of the algorithms, datasets, and specific experimental setups used in the study. If the code and data are publicly available, reproducing the results could be straightforward for those with expertise in machine learning and access to suitable computing resources. However, if the algorithms are proprietary or if the

dataset used is not publicly accessible, reproducing the results could be challenging or impossible.

5. Effort Required for Replication:

- The experimental design and methodologies are described in detail, including performance metrics, sub-analyses, and comparisons with prior studies. This level of detail provides a basis for replication efforts. However, replicating the study exactly might require access to similar datasets, algorithms, and resources, which could pose challenges depending on availability and access to such resources.

Development and Validation of an Artificial Intelligence–Powered Platform for Prostate Cancer Grading and Quantification

Answers to Design Questions given in the Rubric:

1. Workflow of the experiment:
   - The experiment involves the development and training of an artificial intelligence (AI) system to assist in prostate cancer grading and quantification. Three experienced urological pathologists manually graded and quantified biopsy slides, and one of them also annotated the slides for training the algorithm. The AI system then independently graded and quantified the biopsy slides, with pathologists reviewing and modifying the results as necessary.

2. Analysis of experiment outcomes:
   - The outcomes of the experiments were analyzed using various methods, including receiver operating characteristic (ROC) analysis to evaluate the AI system's ability to distinguish between prostate cancer and benign prostatic epithelium and stroma. Linear-weighted Cohen κ values were used to measure concordance between the AI system and the training pathologist, as well as between the three pathologists in prostate cancer grading and quantification. Statistical significance was assessed using appropriate tests such as χ2 tests.

3. Patient and doctor selection:
   - Patients were chosen based on biopsy-confirmed prostate cancer cases within a specific time frame at the University of Wisconsin Health System. Doctors involved in the study were three experienced academic urological pathologists selected for their expertise in grading and quantifying prostate cancer biopsy slides.

4. Proposed benefit of the hybrid system:
   - The proposed benefit of the hybrid system is to improve the efficiency and accuracy of prostate cancer grading and quantification by combining the expertise of human pathologists with the capabilities of an AI-powered platform. This system aims to provide more consistent and reliable results compared to manual grading alone.

5. Short-term and long-term effects on doctors:
   - The short-term effect on doctors is an immediate increase in workload due to the need to review and modify the AI-generated results. However, in the long term, the hybrid system may lead to improved efficiency as pathologists become more familiar with the AI tool and its capabilities. Additionally, it may enhance diagnostic accuracy and reduce fatigue associated with manual grading tasks.

Answers to the Claims Questions given in the Rubric

1. Main claims:
   - The main claims of the study are:

- An AI-enabled platform using deep convolutional neural network models and hybrid architectures can accurately distinguish prostate cancer epithelium from benign glands or epithelium and stroma.
- The platform can identify different Gleason patterns with high accuracy at the patch-pixel level, achieving an area under the curve (AUC) of 0.92.
- The platform can provide instant and precise tabulation of the volume of total prostate cancer and the percentage of Gleason patterns in all cores per biopsy site or slide, which is not achievable through manual analysis.

2. Real-world or specific dataset:
   - The claim is about the real-world application of an AI-enabled platform for prostate cancer grading and quantification. The study validates the platform's effectiveness using a specific dataset but aims for broader applicability in clinical settings.

3. Clarity of the claim in title and abstract:
   - Yes, the claim is expressed clearly in both the title and the abstract. The title mentions an "AI-Powered Platform for Prostate Cancer Grading and Quantification," indicating the focus of the study. The abstract provides a concise overview of the study's findings, including the capabilities of the AI platform.

4. Effort required to reproduce:
   - Reproducing the study would require access to the code and data used for training the AI models, as well as the classifier used in the experiment. The study mentions that the AI platform was developed using Python and Theano deep learning framework, and the models were trained on an Amazon Web Services instance. The availability of the code and data, as well as the classifier, would determine the effort required for reproduction.

5. Effort required to replicate experimental design:
   - The experimental design is given in sufficient detail for replication. The study outlines the process of training the AI models, validating the platform, and comparing the results with manual grading by pathologists. Details about data selection, annotation, model architecture, training process, validation methods, and statistical analysis are provided, enabling replication by researchers in the field. <This message was edited>

Answers to the Evidence Questions given in the Rubric:

1. The size of the data used for evaluation (test data) is mentioned as 162 biopsies in the validation set.

2. The data collection process appears to be sufficiently diverse to support the claim. The study selected a total of 589 men with biopsy-confirmed prostate cancer who received care in the University of Wisconsin Health System over a significant period. The selection criteria aimed to include a representative sample of patients with prostate cancer, and the biopsies were performed using a systematic method. Additionally, the validation set of 162 biopsies was chosen to include an adequate mixture of all Gleason patterns and prostate cancer Gleason

grade groups. This diversity in the data collection process ensures that the AI platform's performance is evaluated across a range of cases, enhancing its generalizability and applicability in clinical settings.