Multi-Cancer Image Classification Using Convolutional Neural Networks

William English and Michael Miller
University of Central Florida

December 6, 2023

Abstract

This project aims to develop a deep-learning model capable of classifying various types of cancer from histopathological images. Given the diverse nature of cancerous cells and the subtleties in their visual presentation, this task presents significant challenges. However, the potential for aiding early diagnosis and treatment planning makes it an important endeavor in medical imaging and cancer research. This report summarizes our approach, methodologies, findings, and the implications of applying convolutional neural networks (CNNs) in the medical field.

1 Introduction

Cancer remains one of the leading causes of mortality worldwide, and early detection is crucial for successful treatment. This project explores the application of convolutional neural networks (CNNs) to classify images of different cancer types, such as Brain, Breast, Cervical, Lymphoma, Kidney, and Oral cancers. The importance of this project lies in its potential to assist pathologists in diagnosing cancer more efficiently and accurately, leveraging the advancements in computer vision and artificial intelligence. The research is relevant to us as Mr. English is a survivor of Hodgkin's Lymphoma, and Mr. Miller's mother is a survivor of Brain and Kidney Cancer. It is a common occurrence in the medical field for someone to have tests performed on them that could reveal cancer before it progresses, but because it was not the original point of the test, this possibility is overlooked. Machine models provide labs or hospitals the ability to passively scan and detect cancer in patient screenings for unrelated issues, such as a pelvic scan for a broken bone revealing early-stage kidney cancer. This idea is still distant in implementation, but our research can promote awareness of its possibility and lead to further improvements.

2 Method

Our approach utilizes a custom-designed CNN model, tailored specifically for the task of multi-cancer classification. The model architecture, as defined in Python using Keras, includes convolutional layers, batch normalization, max pooling, dropout layers, and dense layers, ending with a softmax activation for multi-class classification. We chose this architecture for its proven efficacy in image recognition tasks and its adaptability to our diverse dataset, which includes different sources of images such as MRIs and pap smears. Data preprocessing steps, including image normalization and augmentation techniques like rotations and flips, are used to enhance the model's ability to generalize from a limited set of images.

3 Experiments

3.1 Dataset

The Multi Cancer dataset consists of 130,000 512x512 histopathological and MRI images, representing eight different types of cancer, sourced from publicly available Kaggle and Figshare repositories. We

discuss the characteristics of each dataset in terms of size, resolution, and diversity, and the preprocessing steps applied to each.

For each classification of cancer, there exist sub-classes for the model to learn further. Each subclass has 5,000 images each. There are (1) 20,000 images of Acute Lymphoblastic Leukemia separated into subclasses of Benign, Early, Pre, and Pro; (2) 15,000 images of Brain Cancer subclassed into Glioma, Meningioma, and Pituitary Tumors; (3) 10,000 images of Breast Cancer subclassed into Benign and Malignant Tumors; (4) 25,000 Cervical Cancer images subclassed into Dyskeratotic, Koilocytotic, Metaplastic, Parabasal, and Superficial-Intermediate; (5) 10,000 Kidney Cancer (X-Ray) images of Normal and Tumor subclasses; (6) 25,000 Lung and Colon Cancer images subclassed into Colon Adenocarcinoma, Colon Benign Tissue, Lung Adenocarcinoma, Lung Benign Tissue, and Lung Squamous Cell Carcinoma; (7) 15,000 Lymphoma images of Chronic Lymphocytic Leukemia, Follicular Lymphoma, and Mantle Cell Lymphoma subclasses; and (8) 10,000 Oral Cancer images subclassified as Normal and Oral Squamous Cell Carcinoma. By using Keras' `ImageDataGenerator` on different folder hierarchies, we are able to first train a model on *all* images, classifying only between the 8 main classes, and then train supplementary models on each of the 8 classes to sub-classify our images.

In our analysis of the dataset and model results when trained on this dataset, there appears to be issues with the MRI images from the Kidney Cancer, as this sample type is very different from the histopathological smears or slices present in the other datasets. This will be described further in Section 4.

3.2 Model Training and Architecture

Our approach involved training multiple convolutional neural network (CNN) classifiers to detect various types of cancer. Each classifier was designed to target a specific cancer type, ensuring specialized detection capabilities. The training process was conducted using TensorFlow and Keras libraries.

Each model was structured as a Sequential model, comprising convolutional layers (Conv2D), batch normalization, max pooling (MaxPooling2D), dropout layers, and fully connected dense layers. The architecture for each classifier was identical, with the final dense layer's neuron count set to the number of classes (cancer types) in the respective dataset. The architectural details are as follows:

- **Convolutional Layers:** Initial convolutional layer with 32 filters of size 3x3, followed by a layer with 64 filters of the same size. Each layer utilized ReLU activation.
- **Batch Normalization:** Applied after each convolutional layer to stabilize and accelerate the training process.
- **Max Pooling:** Pooling layers with a 2x2 window were used following each batch normalization step to reduce dimensionality.
- **Dropout:** Dropout layers were employed with a rate of 0.2 after max pooling and 0.4 before the final dense layer to prevent overfitting.
- **Flatten and Dense Layers:** A flattening layer transitioned the data from 2D to 1D, followed by a dense layer with 256 neurons and a final dense output layer with neurons corresponding to the number of classes, using softmax activation.

The models were compiled using the Adam optimizer and an initial learning rate of 0.001. Categorical cross-entropy was chosen as the loss function.

3.3 Data Preparation and Training Process

The training process was executed individually for each cancer type. A unique model instance was created for each type, with the number of output neurons in the final dense layer set according to the number of classes in the respective dataset.

- **Data Preparation:** Training, validation, and test data generators were created for each cancer type, employing data augmentation for robustness.

- **Callbacks:** Early stopping was employed to halt training if validation loss did not improve for 10 epochs, and learning rate reduction on plateau was used, reducing the learning rate by a factor of 0.2 if no improvement in validation loss was observed for 5 epochs.
- **Model Training:** Each model was trained for up to 50 epochs with batch size 16, using the respective training and validation data generators.
- **Model Evaluation:** Post-training, each model was evaluated on a separate test set to assess its performance, measuring both loss and accuracy, as well as generating confusion matrices for each classifier and relevant test set.

The training process was automated to sequentially train a model for each cancer type, ensuring efficiency and consistency across all models. This process took approximately 14 hours in total, training on a single Nvidia A100 GPU with 80 GB of V-ram.

3.4 Evaluation Metrics and results

We report the loss, accuracy, precision, recall, F1-score, and confusion matrices for each classifier we trained.

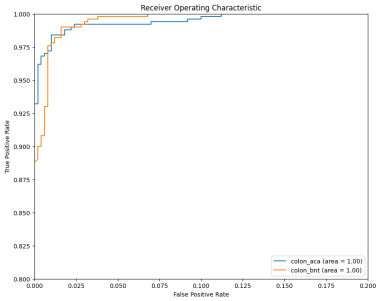
Table 1: Performance Evaluation of Trained Classifiers

Cancer Type	**Loss**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Colon Cancer	0.1683	95.00%	95%	95%	95%
Lung Cancer	0.1667	92.07%	92%	92%	92%
Cervical Cancer	0.1599	94.40%	95%	94%	94%
Breast Cancer	0.1371	94.30%	94%	94%	94%
Lymphoma	0.9008	84.80%	87%	85%	85%
Brain Cancer	0.6191	76.87%	82%	77%	74%

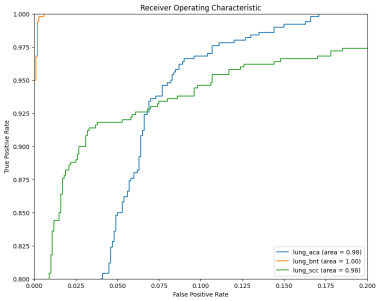
Below, we present ROC Curves related to our results. Please note that these figures are zoomed into the top-left corners of the curves. Because of the zoom level, curves appear "chunky" as the data points become less continuous due to zoom. Still, it is evident that for the different cancer types, certain subclasses are harder to train Precision for. Specifically, malignant labels have a common struggle with keeping a high true positive rate when compared to the benign labels, which optimize well at high TP and low FP rates compared to benign examples. For example, in the Cervical Cancer data set, the model had issues in correctly reporting Metaplastic and Dyskeratotic classes correctly, and had issues with false positives.

Additionally, we present confusion matrices. Analysis of these will be explored below in Section 4: Discussion, as we bridge from the information present in the ROC Curves to the Recall abilities of our model.

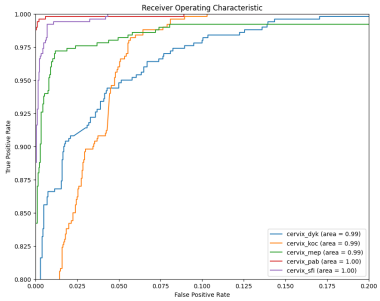
Figure 1: ROC Curves for Various Cancers



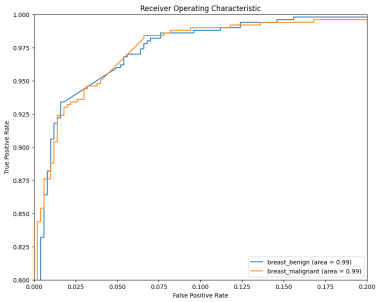
(a) ROC Curve for Colon Cancer



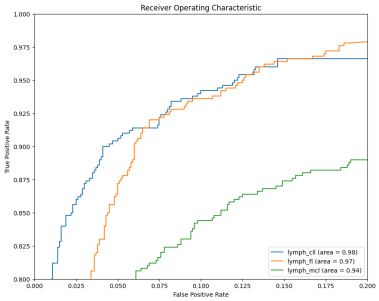
(b) ROC Curve for Lung Cancer



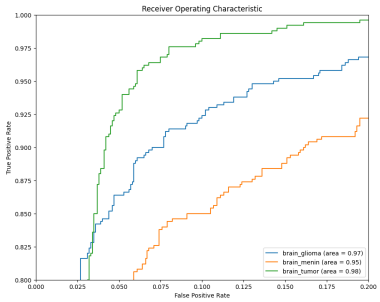
(c) ROC Curve for Cervical Cancer



(d) ROC Curve for Breast Cancer

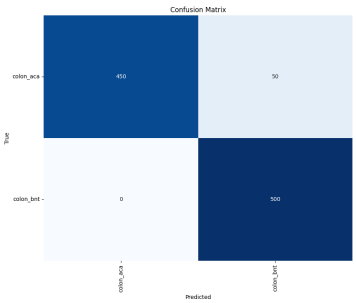


(e) ROC Curve for Lymphoma

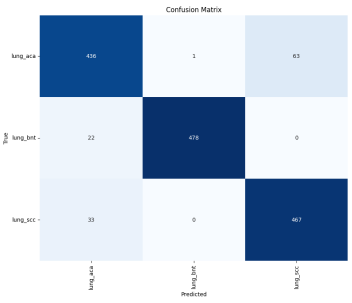


(f) ROC Curve for Brain Cancer

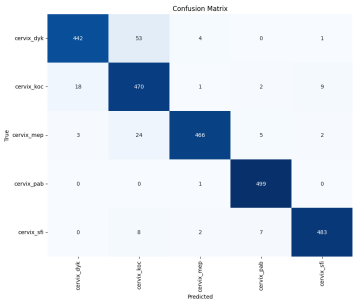
Figure 2: Confusion Matrices for Various Cancers



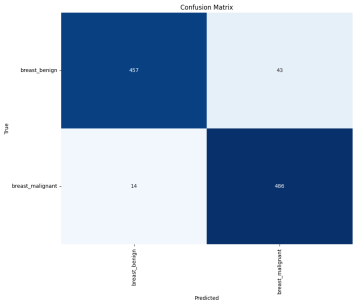
(a) Confusion Matrix for Colon Cancer



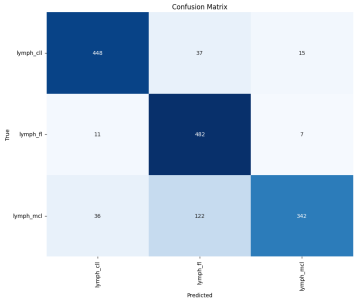
(b) Confusion Matrix for Lung Cancer



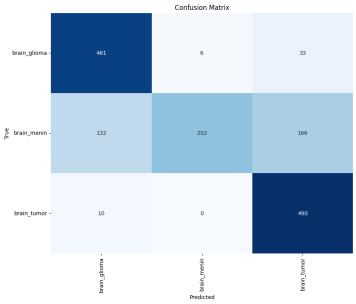
(c) Confusion Matrix for Cervical Cancer



(d) Confusion Matrix for Breast Cancer



(e) Confusion Matrix for Lymphoma



(f) Confusion Matrix for Brain Cancer

4 Discussion

The application of our work, as well as any work in medical image analysis through deep learning, should be performed with heavy moderation and careful tact. From a numbers standpoint, our model performs extremely well with a high accuracy score for many of the classifications of cancer trained upon, as evident through our confusion matrices. However, we must consider the potential social impact of our work when discussing these numbers: each training instance (and in a production environment, each novel instance predicted by the model) is a real human life whose life can be altered by the results of a diagnosis. The class predictions from a CV model should only be supplementary to a medical professional's work, and could never replace it, but may help medical personnel prioritize more serious degrees of cancer progression since a computer vision model can quicker identify severity than a doctor could.

In cases like the Brain Cancer, Cervical Cancer, and Lymphoma datasets, the presumption is that the patient has already been diagnosed with cancer, however, assistance is needed in specifying the diagnosis from there. Meningioma seemed to confuse the model quite a lot, and this is partly due to the similarity between it and the other two Brain Cancer types present in the dataset, Glioma and Pituitary Tumors, whereas the other two have more differences between each other. It is also due to the fact that while it is easier to differentiate between these types of cancer when referencing something like an MRI or CT Scan image, histological examination is tougher to perform between these types without being a trained medical professional.

Cervical Cancer results were very promising, with only a few false predictions between classes. However, discussion of this category helps us approach a bigger concern we have not explicitly described yet. Dyskeratotis means the irreversible degeneration of skin cells, specifically prematurely. In the context of a cervix, dyskeratotic cervical cells imply something is wrong with the patient, but while largely cited to relate to pre-cancerous cell changes, it could also be the result of a Human Papillomavirus (HPV) infection or something else. This helps explain why our model had high confusion between dyskeratosis and koilocytotic classifications, because Koilocytosis means you have koilocytes, cells that develop after you've had a human papillomavirus infection. It is an important difference to make in classification, though, because Koilocytosis is the most common cytopathic effect present in HPV-related oropharyngeal cancers. Metaplastic and Parabasal results in a pap smear suggest the patient has cervical cancer, with parabasal cells meaning a very high chance of cancer, and Superficial-Intermediate means they have cancer. The model performs really well with the latter two classifications, thankfully, because a false negative or false positive with these could lead to the loss of human life or the devastation of one's psyche if they are misdiagnosed. The one misclassification of dyskeratotic cells as superficial-intermediate would drastically harm one's life if medical staff worked only off of the assumption that the AI model is to be trusted with high trust. Even worse, the eight misclassifications of Superficial-Intermediate cancer as koilocytotic cells could kill a patient over time if staff were to trust the predictions, as the severity of their condition would go ignored. This is the reality we must reconcile with when analyzing the implications and limitations of the datasets we train on for something like medical analysis: With no (largely-populated) basis of healthy individuals in this Cervical Cancer dataset, as well as in the Brain Cancer one, it is important that this work be used only as a supplement to a medical assumption. The model can help guide nurses or others to seek the right professionals for diagnosis, but only these professionals should make the final call, using more data than just our predictions, as the "rare" instance of a misclassification would cost a human life if trusted as ground truth 100% of the time.

Thankfully, we are able to shift our focus to datasets with large populations of healthier individuals than we just analyzed. Breast Cancer, Colon Cancer, and Lung Cancer datasets all had training instances of benign tissue. While this is not as useful as the "Normal" labels present in the Kidney Cancer and Oral Cancer datasets we did not use in this report, they are a valuable step in the right direction as "Benign" implies the presence of cancer cells already. Therefore, the results of these models should only be given trust if medical staff already know the patient in question has cancerous cells they are taking samples of. Our model results for colon cancer thankfully did not predict any benign cells as Adenocarcinoma, but it did falsely classify 50 instances of Colon Adenocarcinoma as benign tissue. In the sphere of medical diagnoses, these false negatives are far more catastrophic in impact than a false positive, because the false positive would eventually be discovered as false whereas a false negative could send a patient home to die without knowing what has ailed them. It is imperative, therefore, for a trained model to make no mistakes of the False Negative variety, as it disturbs the

delicacy of life many degrees more than a False Positive does. Future iterations of this model would need a loss function that heavily penalizes False Negatives, as described later on. Benign Lung tissue was occasionally falsely flagged as Adenocarcinoma, but never as Squamous Cell Carcinoma. Through some research into where the distinctions are between these histopathologically, it is clear that the benign lung cells sampled are similar in appearance to the Adenocarcinoma cells. Squamous cells are flatter, scale-like cells with a very different appearance, so this is likely what was learned to some degree. It appears that Adenocarcinoma is doubly more likely to be misclassified as Squamous Cell Carcinoma than the reverse event. Thankfully, in our testing battery our model only gave one False Negative flag to an Adenocarcinoma sample. This is still a grave mistake, in our opinion, but it is better than the other models we have discussed. A larger dataset with more verbosely described sampling techniques would be necessary for future experiments, as would potentially different machine learning frameworks. While our model is not perfect in correctly diagnosing the two malignant cancers it trained on, it can identify benign cancers with reasonably high confidence. Malignant breast tissue had 14 False Negative events in our testing battery, which is too low a score for us to prescribe as a good result. Our model, again, could help hospitals prioritize malignant patients in the queue of receiving care, but it is imperative that there is still a team of professionals overseeing the backlog of "benign" predictions, as False Negative results like these can result in false relief and loss of life.

Overall, most of these results are promising, considering our computationally inexpensive model was trained only for 50 epochs on each dataset. Students of computer vision may find our Accuracy, Precision, Recall, and F1 scores to be impressive, but students of medicine are likely to disagree, considering each negative tick on Recall implies a False Negative event, which is a potentially fatal mistake for patients. The Multi-Cancer dataset provides an excellent foray into the analysis of open medical data for improving human life but has its flaws in terms of what a model can train on. Since this dataset's labels are not directly applicable to segmentation tasks, it is possible that many training images have cells present of both the cancerous and non-cancerous variety. Unless a dataset was designed that only shows one cell (in the scope of histopathological samples) or segment of the body (in the scope of the Brain and Kidney datasets' scans) at a time, a model may have issues in choosing between two classifications if an image contains both good and bad results. In future models, "bad" (cancer) results should be the prioritized report, as high Recall (False-Negative reduction) is more important in medicine than the optimization of Precision (False-Positive reduction), since false positive predictions would be discovered by treatment professionals rather quickly. Additionally, most of the data in the Multi-Cancer dataset has been sourced from different countries such as Bangladesh and Iran. There is nothing inherently wrong with this, as it is critical to sample many different parts of the world, but a real-world application of our model would need to be trained on either (1) a supermassive dataset sampling from many populations worldwide; or (2) a dataset sampling from the area the model is being used, such as a model used in American hospitals being trained on North American, South American, Eastern Asian, and Western European datasets, or a model used in Azerbaijani hospitals being trained on Western European, Western Asian, North African, and Middle Eastern datasets.

Through some ERD, it is also evident that many of the training samples are imperfect. Blurriness, chromatic aberrations, and cropping techniques that stretched sample images to meet resolution demands haunt many of the images. Additionally, data sets like Brain Cancer contain scans of the brain from different angles and some are mirrored. While this could be good for training from a vision perspective, we propose the idea of a standardized input orientation for the scans such that models are only learning *what cancer looks like*, rather than attributing classification to the orientation of the image. That is a possibility if there were, for instance, more training samples of Glioma with the head facing rightward, and more benign samples facing leftward, as a benign sample facing rightward could be misclassified. Of course, that is why we train with random rotations, flips, and crops, but the dataset may still be limiting the efficiency of our trained model due to these statistical distributions. Alternative approaches to improving the model would include a more standardized dataset with fewer of these quirks between samples.

5 Conclusion

5.1 Future Work

It is imperative that in the sphere of medical diagnoses, a Machine model's **confidence levels in any prediction should be scrutinized heavily** during and after training, no matter the architecture used or the dataset being explored. The confidence levels of a model's prediction are available during training, and due to the sensitivity of the impacts a prediction can have on one's life, these levels should be analyzed with low-confidence predictions being flagged for review or discarded completely. A hands-on approach of review would be to let the model train as usual, but take low-confidence outputs with knowledge of the confidence, handing the control over to medical professionals to make the diagnosis. A hands-off approach of review would be implemented during training: low-confidence scores are to be penalized during training, possibly through a loss function or a different penalty mechanic as seen in Genetic Ensemble Learning or in the field of Evolutionary Computation. As introduced in our Discussion, these future models should **heavily penalize false negatives** as they present the most risk of a life being damaged by a computer's prediction.

In future research into this application of computer vision technology, learning architectures should aim to implement not just convolutional techniques, but also techniques such as Vision Transformers, Graph Analysis, and instance segmentation while training. Instance segmentation is useful for separating the cells present in a histopathological image, as well as the organs present in an MRI or other imaging scan, such that "non-cancerous" and "cancerous" predictions could be made (with further subclassifications from there, such as "at-risk" cells in the case of Dyskeratosis, or the type of cancer present in the cases of Glioma, Pituitary Tumors, or Meningioma). In review of the data we trained with, we note that it is also critical for datasets to have many more "normal" instances. Previously, we described identifying benign cancers with a reasonably high confidence level. We present the idea of *"Low-Confidence Suppression"* in future research, with a focus on false negative events being penalized heavily during training.

Additionally, there are a few types of cancer included in the Multi-Cancer dataset for which we did not train classifiers on, for various reasons. These include Oral Cancer, Kidney Cancer, and Acute Lymphoblastic Leukemia (ALL). We believe our method of training classifiers would be comparably effective on the Oral Cancer and ALL datasets, but the Kidney Cancer dataset will require a new approach, as the data is not hisopathological cell images, but rather MRI images from multiple angles. To prove this, we kept the Brain Cancer dataset in our training. These samples are of T1-weighted contrast MRI scans, and our Brain model was the lowest-performing (76.87% accuracy), showing that the Kidney Cancer analysis is possible, but due to the disparity between tissue imaging and histopathological imaging, these sets should see different approaches taken. Further work could also be done to improve the ability of the Brain Cancer classifier to distinguish Meningioma from Glioma and Pituitary Tumors, as well as to improve the ability of the Lymphoma classifier to distinguish Follicular Lymphoma from Mantle Cell Lymphoma.

5.2 Code

The code for our model is hosted on this GitHub repository. A brief description of the code structure and instructions for running the model are provided.

5.3 Contributions

William was responsible for building and training the classifiers, as well as implementation of the evaluation code. The HPC used to train our models was provided by the group William works for at UCF, the AI and Emerging Computing Lab.

Michael was responsible for the medical research present throughout this document, as well as validating the training done by William on UCF's Newton HPC. The results were of similar statistics, so they were not introduced any further in the paper.

6 References

Naren, O. (2023). Multi Cancer. Kaggle. Retrieved from https://www.kaggle.com/datasets/obulisainaren/multi-cancer

Boras, VF, and MA Duggan. "Cervical Dyskeratotic Cells as Predictors of Condylomatous Changes on Biopsy." PubMed, Published by the Department of Pathology, Foothills Hospital, Calgary, Alberta, Canada. Publicly hosted by the U.S. National Library of Medicine, 1989, pubmed.ncbi.nlm.nih.gov/2538985/.

Wang, Wen, Hui Liu, and Guoli Li. "What's the difference between lung adenocarcinoma and lung squamous cell carcinoma? Evidence from a retrospective analysis in a cohort of Chinese patients" PubMed Central, article was submitted to Cancer Endocrinology, a section of the journal Frontiers in Endocrinology Publicly hosted by the U.S. National Library of Medicine, 1989, https://www.ncbi.nlm.nih.gov/pmc/articles

Abadi, M., Agarwal, A., Barham, P., et al. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.

Bradski, G., & Kaehler, A. (2008). Learning OpenCV: Computer vision with the OpenCV library. O'Reilly Media, Inc.

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Rafal Jozefowicz, Yangqing Jia, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Mike Schuster, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.