Predicting cancer outcomes from histology and genomics using convolutional networks

Pooya Mobadersany[a], Safoora Yousefi[a], Mohamed Amgad[a], David A. Gutman[b], Jill S. Barnholtz-Sloan[c], José E. Velázquez Vega[d], Daniel J. Brat[e], and Lee A. D. Cooper[a,f,g,1]

[a]Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA 30322; [b]Department of Neurology, Emory University School of Medicine, Atlanta, GA 30322; [c]Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, Cleveland, OH 44106; [d]Department of Pathology and Laboratory Medicine, Emory University School of Medicine, Atlanta, GA 30322; [e]Department of Pathology, Northwestern University Feinberg School of Medicine, Chicago, IL 60611; [f]Winship Cancer Institute, Emory University, Atlanta, GA 30322; and [g]Department of Biomedical Engineering, Emory University and Georgia Institute of Technology, Atlanta, GA 30322

Edited by Bert Vogelstein, Johns Hopkins University, Baltimore, MD, and approved February 13, 2018 (received for review October 4, 2017)

Cancer histology reflects underlying molecular processes and disease progression and contains rich phenotypic information that is predictive of patient outcomes. In this study, we show a computational approach for learning patient outcomes from digital pathology images using deep learning to combine the power of adaptive machine learning algorithms with traditional survival models. We illustrate how these survival convolutional neural networks (SCNNs) can integrate information from both histology images and genomic biomarkers into a single unified framework to predict time-to-event outcomes and show prediction accuracy that surpasses the current clinical paradigm for predicting the overall survival of patients diagnosed with glioma. We use statistical sampling techniques to address challenges in learning survival from histology images, including tumor heterogeneity and the need for large training cohorts. We also provide insights into the prediction mechanisms of SCNNs, using heat map visualization to show that SCNNs recognize important structures, like microvascular proliferation, that are related to prognosis and that are used by pathologists in grading. These results highlight the emerging role of deep learning in precision medicine and suggest an expanding utility for computational analysis of histology in the future practice of pathology.

artificial intelligence | machine learning | digital pathology | deep learning | cancer

Histology has been an important tool in cancer diagnosis and prognostication for more than a century. Anatomic pathologists evaluate histology for characteristics, like nuclear atypia, mitotic activity, cellular density, and tissue architecture, incorporating cytologic details and higher-order patterns to classify and grade lesions. Although prognostication increasingly relies on genomic biomarkers that measure genetic alterations, gene expression, and epigenetic modifications, histology remains an important tool in predicting the future course of a patient's disease. The phenotypic information present in histology reflects the aggregate effect of molecular alterations on cancer cell behavior and provides a convenient visual readout of disease aggressiveness. However, human assessments of histology are highly subjective and are not repeatable; hence, computational analysis of histology imaging has received significant attention. Aided by advances in slide scanning microscopes and computing, a number of image analysis algorithms have been developed for grading (1–4), classification (5–10), and identification of lymph node metastases (11) in multiple cancer types.

Deep convolutional neural networks (CNNs) have emerged as an important image analysis tool and have shattered performance benchmarks in many challenging applications (12). The ability of CNNs to learn predictive features from raw image data is a paradigm shift that presents exciting opportunities in medical imaging (13–15). Medical image analysis applications have heavily relied on feature engineering approaches, where algorithm pipelines are used to explicitly delineate structures of interest using segmentation algorithms to measure predefined features of these structures that are believed to be predictive and to use these features to train models that predict patient outcomes. In contrast, the feature learning paradigm of CNNs adaptively learns to transform images into highly predictive features for a specific learning objective. The images and patient labels are presented to a network composed of interconnected layers of convolutional filters that highlight important patterns in the images, and the filters and other parameters of this network are mathematically adapted to minimize prediction error. Feature learning avoids biased a priori definition of features and does not require the use of segmentation algorithms that are often confounded by artifacts and natural variations in image color and intensity. While feature learning has become the dominant paradigm in general image analysis tasks, medical applications pose unique challenges. Large amounts of labeled data are needed to train CNNs, and medical applications often suffer from data deficits that limit performance. As "black box" models, CNNs are also difficult to deconstruct, and therefore, their prediction mechanisms are difficult to interpret. Despite these

Significance

Predicting the expected outcome of patients diagnosed with cancer is a critical step in treatment. Advances in genomic and imaging technologies provide physicians with vast amounts of data, yet prognostication remains largely subjective, leading to suboptimal clinical management. We developed a computational approach based on deep learning to predict the overall survival of patients diagnosed with brain tumors from microscopic images of tissue biopsies and genomic biomarkers. This method uses adaptive feedback to simultaneously learn the visual patterns and molecular biomarkers associated with patient outcomes. Our approach surpasses the prognostic accuracy of human experts using the current clinical standard for classifying brain tumors and presents an innovative approach for objective, accurate, and integrated prediction of patient outcomes.

Author contributions: P.M., S.Y., M.A., D.A.G., D.J.B., and L.A.D.C. designed research; P.M., S.Y., J.E.V.V., and L.A.D.C. performed research; P.M., J.S.B.-S., and L.A.D.C. analyzed data; and P.M., M.A., D.A.G., J.S.B.-S., J.E.V.V., D.J.B., and L.A.D.C. wrote the paper.

Conflict of interest statement: L.A.D.C. leads a research project that is financially supported by Ventana Medical Systems, Inc. While this project is not directly related to the manuscript, it is in the general area of digital pathology.

This article is a PNAS Direct Submission.

Data deposition: Software and other resources related to this paper have been deposited at GitHub, https://github.com/CancerDataScience/SCNN.

[1]To whom correspondence should be addressed. Email: Lee.Cooper@Emory.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1717139115/-/DCSupplemental.

Published online March 12, 2018.

challenges, CNNs have been successfully used extensively for medical image analysis (9, 11, 16–26).

Many important problems in the clinical management of cancer involve time-to-event prediction, including accurate prediction of overall survival and time to progression. Despite overwhelming success in other applications, deep learning has not been widely applied to these problems. Survival analysis has often been approached as a binary classification problem by predicting dichotomized outcomes at a specific time point (e.g., 5-y survival) (27). The classification approach has important limitations, as subjects with incomplete follow-up cannot be used in training, and binary classifiers do not model the probability of survival at other times. Time-to-event models, like Cox regression, can utilize all subjects in training and model their survival probabilities for a range of times with a single model. Neural network-based Cox regression approaches were explored in early machine learning work using datasets containing tens of features, but subsequent analysis found no improvement over basic linear Cox regression (28). More advanced "deep" neural networks that are composed of many layers were recently adapted to optimize Cox proportional hazard likelihood and were shown to have equal or superior performance in predicting survival using genomic profiles containing hundreds to tens of thousands of features (29, 30) and using basic clinical profiles containing 14 features (31).

Learning survival from histology is considerably more difficult, and a similar approach that combined Cox regression with CNNs to predict survival from lung cancer histology achieved only marginally better than random accuracy (0.629 c index) (32). Time-to-event prediction faces many of the same challenges as other applications where CNNs are used to analyze histology. Compared with genomic or clinical datasets, where features have intrinsic meaning, a "feature" in an image is a pixel with meaning that depends entirely on context. Convolution operations can learn these contexts, but the resulting networks are complex, often containing more than 100 million free parameters, and thus, large cohorts are needed for training. This problem is intensified in time-to-event prediction, as clinical follow-up is often difficult to obtain for large cohorts. Data augmentation techniques have been adopted to address this problem, where randomized rotations and transformations of contrast and brightness are used to synthesize additional training data (9, 11, 14, 15, 17, 19, 25, 26, 33). Intratumoral heterogeneity also presents a significant challenge in time-to-event prediction, as a tissue biopsy often contains a range of histologic patterns that correspond to varying degrees of disease progression or aggressiveness. The method for integrating information from heterogeneous regions within a sample is an important consideration in predicting outcomes. Furthermore, risk is often reflected in subtle changes in multiple histologic criteria that can require years of specialized training for human pathologists to recognize and interpret. Developing an algorithm that can learn the continuum of risks associated with histology can be more challenging than for other learning tasks, like cell or region classification.

In this paper, we present an approach called survival convolutional neural networks (SCNNs), which provide highly accurate prediction of time-to-event outcomes from histology images. Using diffuse gliomas as a driving application, we show how the predictive accuracy of SCNNs is comparable with manual histologic grading by neuropathologists. We further extended this approach to integrate both histology images and genomic biomarkers into a unified prediction framework that surpasses the prognostic accuracy of the current WHO paradigm based on genomic classification and histologic grading. Our SCNN framework uses an image sampling and risk filtering technique that significantly improves prediction accuracy by mitigating the effects of intratumoral heterogeneity and deficits in the availability of labeled data for training. Finally, we use heat map visualization techniques applied to whole-slide images to show how SCNNs learn to recognize important histologic structures that neuropathologists use in grading diffuse gliomas and suggest relevance for patterns with prognostic significance that is not currently appreciated. We systematically validate our approaches by predicting overall survival in gliomas using data from The Cancer Genome Atlas (TCGA) Lower-Grade Glioma (LGG) and Glioblastoma (GBM) projects.

Results

Learning Patient Outcomes with Deep Survival Convolutional Neural Networks. The SCNN model architecture is depicted in Fig. 1 (Fig. S1 shows a detailed diagram). H&E-stained tissue sections are first digitized to whole-slide images. These images are reviewed using a web-based platform to identify regions of interest (ROIs) that contain viable tumor with representative histologic characteristics and that are free of artifacts (*Methods*) (34, 35). High-power fields (HPFs) from these ROIs are then used to train a deep convolutional network that is seamlessly integrated with a Cox proportional hazards model to predict patient outcomes. The network is composed of interconnected layers of image processing operations and nonlinear functions that sequentially transform the HPF image into highly predictive prognostic features. Convolutional layers first extract visual features from the HPF at multiple scales using convolutional kernels and pooling operations. These image-derived features feed into fully connected layers that perform additional transformations, and then, a final Cox model layer outputs a prediction of patient risk. The interconnection weights and convolutional kernels are trained by comparing risk predicted by the network with survival or other time-to-event outcomes using a backpropagation technique to optimize the statistical likelihood of the network (*Methods*).

To improve the performance of SCNN models, we developed a sampling and risk filtering technique to address intratumoral heterogeneity and the limited availability of training samples (Fig. 2). In training, new HPFs are randomly sampled from each ROI at the start of each training iteration, providing the SCNN model with a fresh look at each patient's histology and capturing heterogeneity within the ROI. Each HPF is processed using standard data augmentation techniques that randomly transform the field to reinforce network robustness to tissue orientation and variations in staining (33). The SCNN is trained using multiple transformed HPFs for each patient (one for each ROI) to further account for intratumoral heterogeneity across ROIs. For prospective prediction, we first sample multiple HPFs within each ROI to generate a representative collection of fields for the patient. The median risk is calculated within each ROI, and then, these median risks are sorted and filtered to predict a robust patient-level risk that reflects the aggressiveness of their disease while rejecting any outlying risk predictions. These sampling and filtering procedures are described in detail in *Methods*.

Assessing the Prognostic Accuracy of SCNN. To assess the prognostic accuracy of SCNN, we assembled whole-slide image tissue sections from formalin-fixed, paraffin-embedded specimens and clinical follow-up for 769 gliomas from the TCGA (Dataset S1). This dataset comprises lower-grade gliomas (WHO grades II and III) and glioblastomas (WHO grade IV), contains both astrocytomas and oligodendrogliomas, and has overall survivals ranging from less than 1 to 14 y or more. A summary of demographics, grades, survival, and molecular subtypes for this cohort is presented in Table S1. The Digital Slide Archive was used to identify ROIs in 1,061 H&E-stained whole-slide images from these tumors.

The prognostic accuracy of SCNN models was assessed using Monte Carlo cross-validation. We randomly split our cohort into paired training (80%) and testing (20%) sets to generate 15 training/testing set pairs. We trained an SCNN model using

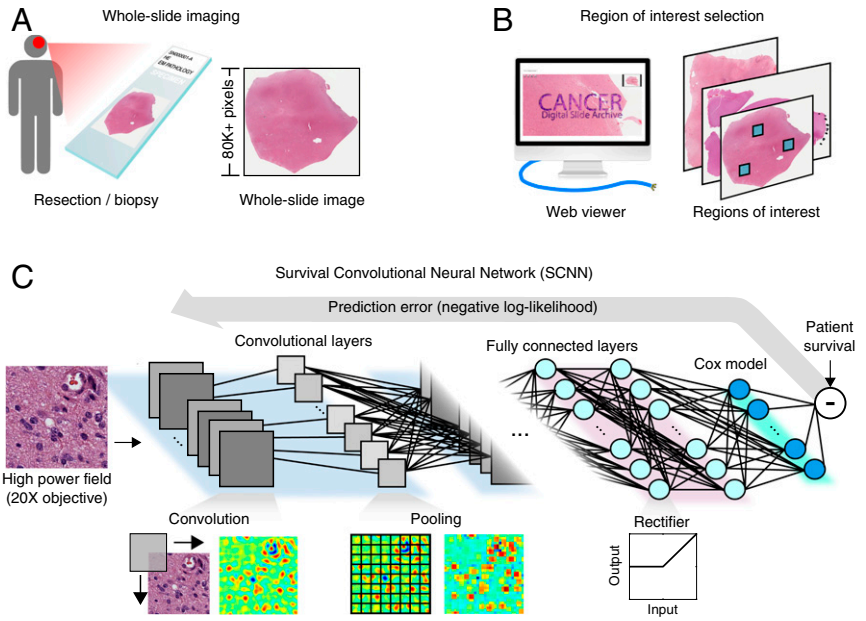

Fig. 1. The SCNN model. The SCNN combines deep learning CNNs with traditional survival models to learn survival-related patterns from histology images. (*A*) Large whole-slide images are generated by digitizing H&E-stained glass slides. (*B*) A web-based viewer is used to manually identify representative ROIs in the image. (*C*) HPFs are sampled from these regions and used to train a neural network to predict patient survival. The SCNN consists of (*i*) convolutional layers that learn visual patterns related to survival using convolution and pooling operations, (*ii*) fully connected layers that provide additional nonlinear transformations of extracted image features, and (*iii*) a Cox proportional hazards layer that models time-to-event data, like overall survival or time to progression. Predictions are compared with patient outcomes to adaptively train the network weights that interconnect the layers.

each training set and then, evaluated the prognostic accuracy of these models on the paired testing sets, generating a total of 15 accuracy measurements (*Methods* and Dataset S1). Accuracy was measured using Harrell's c index, a nonparametric statistic that measures concordance between predicted risks and actual survival (36). A c index of 1 indicates perfect concordance between predicted risk and overall survival, and a c index of 0.5 corresponds to random concordance.

For comparison, we also assessed the prognostic accuracy of baseline linear Cox models generated using the genomic biomarkers and manual histologic grades from the WHO classification of gliomas (Fig. 3*A*). The WHO assigns the diffuse gliomas

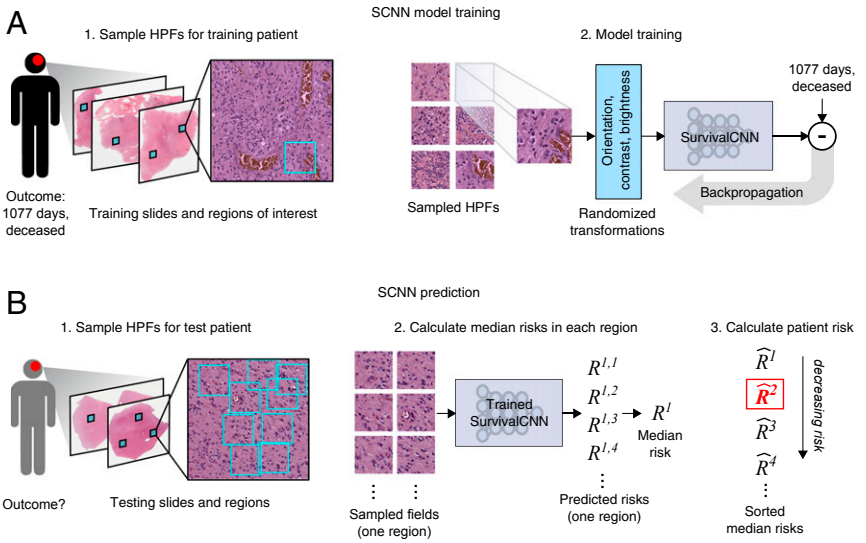

Fig. 2. SCNN uses image sampling and filtering to improve the robustness of training and prediction. (*A*) During training, a single 256 × 256-pixel HPF is sampled from each region, producing multiple HPFs per patient. Each HPF is subjected to a series of random transformations and is then used as an independent sample to update the network weights. New HPFs are sampled at each training epoch (one training pass through all patients). (*B*) When predicting the outcome of a newly diagnosed patient, nine HPFs are sampled from each ROI, and a risk is predicted for each field. The median HPF risk is calculated in each region, these median risks are then sorted, and the second highest value is selected as the patient risk. This sampling and filtering framework was designed to deal with tissue heterogeneity by emulating manual histologic evaluation, where prognostication is typically based on the most malignant region observed within a heterogeneous sample. Predictions based on the highest risk and the second highest risk had equal performance on average in our experiments, but the maximum risk produced some outliers with poor prediction accuracy.

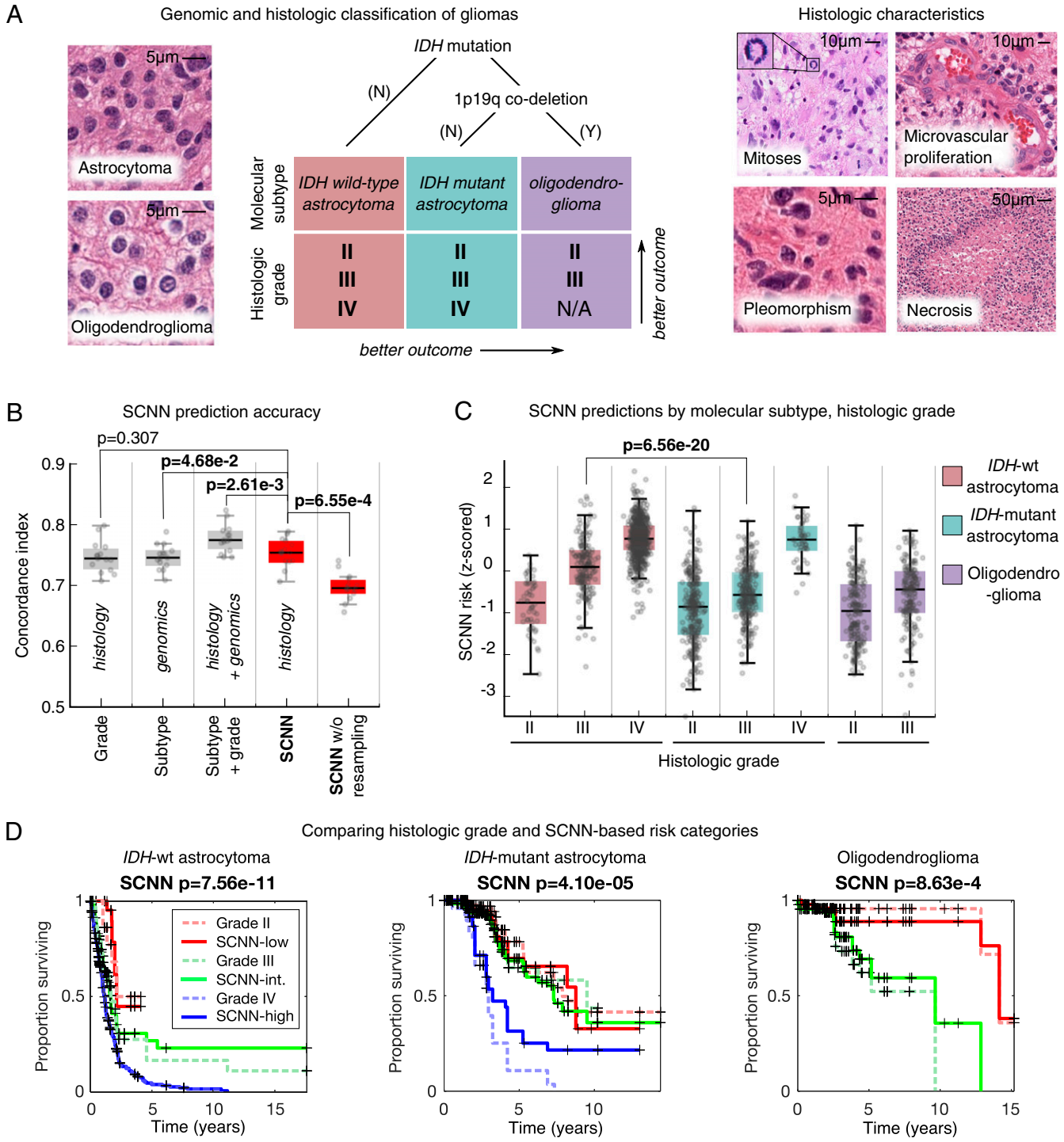

Fig. 3. Prognostication criteria for diffuse gliomas. (*A*) Prognosis in the diffuse gliomas is determined by genomic classification and manual histologic grading. Diffuse gliomas are first classified into one of three molecular subtypes based on *IDH1*/*IDH2* mutations and the codeletion of chromosomes 1p and 19q. Grade is then determined within each subtype using histologic characteristics. Subtypes with an astrocytic lineage are split by *IDH* mutation status, and the combination of 1p/19q codeletion and *IDH* mutation defines an oligodendroglioma. These lineages have histologic differences; however, histologic evaluation is not a reliable predictor of molecular subtype (37). Histologic criteria used for grading range from nuclear morphology to higher-level patterns, like necrosis or the presence of abnormal microvascular structures. (*B*) Comparison of the prognostic accuracy of SCNN models with that of baseline models based on molecular subtype or molecular subtype and histologic grade. Models were evaluated over 15 independent training/testing sets with randomized patient assignments and with/without training and testing sampling. (*C*) The risks predicted by the SCNN models correlate with both histologic grade and molecular subtype, decreasing with grade and generally trending with the clinical aggressiveness of genomic subtypes. (*D*) Kaplan–Meier plots comparing manual histologic grading and SCNN predictions. Risk categories (low, intermediate, high) were generated by thresholding SCNN risks. N/A, not applicable.

to three genomic subtypes defined by mutations in the isocitrate dehydrogenase (*IDH*) genes (*IDH1*/*IDH2*) and codeletion of chromosomes 1p and 19q. Within these molecular subtypes, gliomas are further assigned a histologic grade based on criteria that vary

depending on cell of origin (either astrocytic or oligodendroglial). These criteria include mitotic activity, nuclear atypia, the presence of necrosis, and the characteristics of microvascular structures (microvascular proliferation). Histologic grade remains a significant determinant in planning treatment for gliomas, with grades III and IV typically being treated aggressively with radiation and concomitant chemotherapy.

SCNN models showed substantial prognostic power, achieving a median c index of 0.754 (Fig. 3*B*). SCNN models also performed comparably with manual histologic-grade baseline models (median c index 0.745, $P = 0.307$) and with molecular subtype baseline models (median c index 0.746, $P = 4.68e\text{-}2$). Baseline models representing WHO classification that integrate both molecular subtype and manual histologic grade performed slightly better than SCNN, with a median c index of 0.774 (Wilcoxon signed rank $P = 2.61e\text{-}3$).

We also evaluated the impact of the sampling and ranking procedures shown in Fig. 2 in improving the performance of SCNN models. Repeating the SCNN experiments without these sampling techniques reduced the median c index of SCNN models to 0.696, significantly worse than for models where sampling was used ($P = 6.55e\text{-}4$).

SCNN Predictions Correlate with Molecular Subtypes and Manual Histologic Grade. To further investigate the relationship between SCNN predictions and the WHO paradigm, we visualized how risks predicted by SCNN are distributed across molecular subtype and histologic grade (Fig. 3*C*). SCNN predictions were highly correlated with both molecular subtype and grade and were consistent with expected patient outcomes. First, within each molecular subtype, the risks predicted by SCNN increase with histologic grade. Second, predicted risks are consistent with the published expected overall survivals associated with molecular subtypes (37). *IDH* WT astrocytomas are, for the most part, highly aggressive, having a median survival of 18 mo, and the collective predicted risks for these patients are higher than for patients from other subtypes. *IDH* mutant astrocytomas are another subtype with considerably better overall survival ranging from 3 to 8 y, and the predicted risks for patients in this subtype are more moderate. Notably, SCNN risks for *IDH* mutant astrocytomas are not well-separated for grades II and III, consistent with reports of histologic grade being an inadequate predictor of outcome in this subtype (38). Infiltrating gliomas with the combination of *IDH* mutations and codeletion of chromosomes 1p/19q are classified as oligodendrogliomas in the current WHO schema, and these have the lowest overall predicted risks consistent with overall survivals of 10+ y (37, 39). Finally, we noted a significant difference in predicted risks when comparing the *IDH* mutant and *IDH* WT grade III astrocytomas (rank sum $P = 6.56e\text{-}20$). These subtypes share an astrocytic lineage and are graded using identical histologic criteria. Although some histologic features are more prevalent in *IDH*-mutant astrocytomas, these features are not highly specific or sensitive to *IDH* mutant tumors and cannot be used to reliably predict IDH mutation status (40). Risks predicted by SCNN are consistent with worse outcomes for *IDH* WT astrocytomas in this case (median survival 1.7 vs. 6.3 y in the *IDH* mutant counterparts), suggesting that SCNN models can detect histologic differences associated with *IDH* mutations in astrocytomas.

We also performed a Kaplan–Meier analysis to compare manual histologic grading with "digital grades" based on SCNN risk predictions (Fig. 3*D*). Low-, intermediate-, and high-risk categories were established by setting thresholds on SCNN predictions to reflect the proportions of manual histologic grades in each molecular subtype (*Methods*). We observed that, within each subtype, the differences in survival captured by SCNN risk categories are highly similar to manual histologic grading. SCNN risk categories and manual histologic grades have similar prognostic power in *IDH* WT astrocytomas (log rank $P = 1.23e\text{-}12$ vs. $P = 7.56e\text{-}11$, respectively). In *IDH* mutant astrocytomas, both SCNN risk categories and manual histologic grades have difficulty separating Kaplan–Meier curves for grades II and III, but both clearly distinguish grade IV as being associated with worse outcomes. Discrimination for oligodendroglioma survival is also similar between SCNN risk categories and manual histologic grades (log rank $P = 9.73e\text{-}7$ vs. $P = 8.63e\text{-}4$, respectively).

Improving Prognostic Accuracy by Integrating Genomic Biomarkers. To integrate both histologic and genomic data into a single unified prediction framework, we developed a genomic survival convolutional neural network (GSCNN model). The GSCNN learns from genomics and histology simultaneously by incorporating genomic data into the fully connected layers of the SCNN (Fig. 4). Both data are presented to the network during training, enabling genomic variables to influence the patterns learned by the SCNN by providing molecular subtype information.

We repeated our experiments using GSCNN models with histology images, *IDH* mutation status, and 1p/19q codeletion as inputs and found that the median c index improved from 0.754 to 0.801. The addition of genomic variables improved performance by 5% on average, and GSCNN models significantly outperform the baseline WHO subtype-grade model trained on equivalent data (signed rank $P = 1.06e\text{-}2$). To assess the value of integrating genomic variables directly into the network during training, we compared GSCNN with a more superficial integration approach, where an SCNN model was first trained using histology images, and then, the risks from this model were combined with *IDH* and 1p/19q variables in a simple three-variable Cox model (Fig. S2). Processing genomic variables in the fully connected layers and including them in training provided a statistically significant benefit; models trained using the superficial approach performed worse than GSCNN models with median c index decreasing to 0.785 (signed rank $P = 4.68e\text{-}2$).

To evaluate the independent prognostic power of risks predicted by SCNN and GSCNN, we performed a multivariable Cox regression analysis (Table 1). In a multivariable regression that included SCNN risks, subtype, grade, age, and sex, SCNN risks had a hazard ratio of 3.05 and were prognostic when correcting for all other features, including manual grade and molecular subtype ($P = 2.71e\text{-}12$). Molecular subtype was also significant in the SCNN multivariable regression model, but histologic grade was not. We also performed a multivariable regression with GSCNN risks and found GSCNN to be significant ($P = 9.69e\text{-}12$) with a hazard ratio of 8.83. In the GSCNN multivariable regression model, molecular subtype was not significant, but histologic grade was marginally significant. We also used Kaplan–Meier analysis to compare risk categories generated from SCNN and GSCNN (Fig. S3). Survival curves for SCNN and GSCNN were very similar when evaluated on the entire cohort. In contrast, their abilities to discriminate survival within molecular subtypes were notably different.

Visualizing Histologic Patterns Associated with Prognosis. Deep learning networks are often criticized for being black box approaches that do not reveal insights into their prediction mechanisms. To investigate the visual patterns that SCNN models associate with poor outcomes, we used heat map visualizations to display the risks predicted by our network in different regions of whole-slide images. Transparent heat map overlays are frequently used for visualization in digital pathology, and in our study, these overlays enable pathologists to correlate the predictions of highly accurate survival models with the underlying histology over the expanse of a whole-slide image. Heat maps were generated using a trained SCNN model to predict the risk for each nonoverlapping HPF in a whole-slide image. The predicted risks were used to generate a color-coded transparent

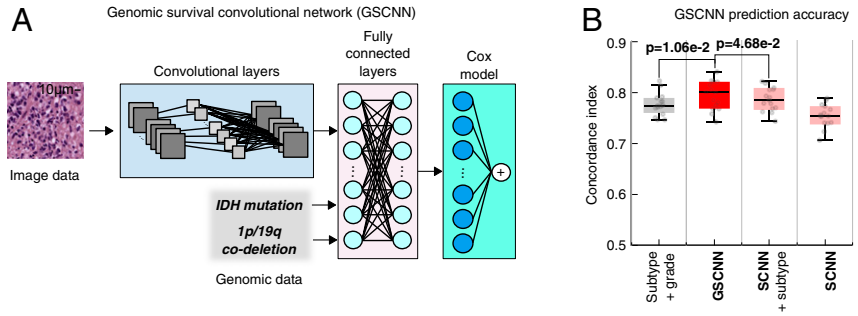

Fig. 4. GSCNN models integrate genomic and imaging data for improved performance. (*A*) A hybrid architecture was developed to combine histology image and genomic data to make integrated predictions of patient survival. These models incorporate genomic variables as inputs to their fully connected layers. Here, we show the incorporation of genomic variables for gliomas; however, any number of genomic or proteomic measurements can be similarly used. (*B*) The GSCNN models significantly outperform SCNN models as well as the WHO paradigm based on genomic subtype and histologic grading.

overlay, where red and blue indicate higher and lower SCNN risk, respectively.

A selection of risk heat maps from three patients is presented in Fig. 5, with inlays showing how SCNNs associate risk with important pathologic phenomena. For TCGA-DB-5273 (WHO grade III, *IDH* mutant astrocytoma), the SCNN heat map clearly and specifically highlights regions of early microvascular proliferation, an advanced form of angiogenesis that is a hallmark of malignant progression, as being associated with high risk. Risk in this heat map also increases with cellularity, heterogeneity in nuclear shape and size (pleomorphism), and the presence of abnormal microvascular structures. Regions in TCGA-S9-A7J0 have varying extents of tumor infiltration ranging from normal brain to sparsely infiltrated adjacent normal regions exhibiting satellitosis (where neoplastic cells cluster around neurons) to moderately and highly infiltrated regions. This heat map correctly associates the lowest risks with normal brain regions and can distinguish normal brain from adjacent regions that are sparsely infiltrated. Interestingly, higher risks are assigned to sparsely infiltrated regions (region 1, *Top*) than to regions containing relatively more tumor infiltration (region 2, *Top*). We observed a similar pattern in TCGA-TM-A84G, where edematous regions (region 1, *Bottom*) adjacent to moderately cellular tumor regions (region 1, *Top*) are also assigned higher risks. These latter examples provide risk features embedded within histologic sections that have been previously unrecognized and could inform and improve pathology practice.

Discussion

We developed a deep learning approach for learning survival directly from histological images and created a unified framework for integrating histology and genomic biomarkers for predicting time-to-event outcomes. We systematically evaluated the prognostic accuracy of our approaches in the context of the current clinical standard based on genomic classification and histologic grading of gliomas. In contrast to a previous study that achieved only marginally better than random prediction accuracy, our approach rivals or exceeds the accuracy of highly trained human experts in predicting survival. Our study provides insights into applications of deep learning in medicine and the integration of histology and genomic data and provides methods for dealing with intratumoral heterogeneity and training data deficits when using deep learning algorithms to predict survival from histology images. Using visualization techniques to gain insights into SCNN prediction mechanisms, we found that SCNNs clearly recognize known and time-honored histologic predictors of poor prognosis and that SCNN predictions suggest prognostic relevance for histologic patterns with significance that is not currently appreciated by neuropathologists.

Our study investigated the ability to predict overall survival in diffuse gliomas, a disease with wide variations in outcomes and an ideal test case where histologic grading and genomic classifications have independent prognostic power. Treatment planning for gliomas is dependent on many factors, including patient age and grade, but gliomas assigned to WHO grades III and IV are typically treated very aggressively with radiation and concomitant chemotherapy, whereas WHO grade II gliomas may be treated with chemotherapy or even followed in some cases (41). Histologic diagnosis and grading of gliomas have been limited by considerable intra- and interobserver variability (42). While the emergence of molecular subtyping has resolved uncertainty related to lineage, criteria for grading need to be redefined in the context of molecular subtyping. For example, some morphologic features used to assess grade (e.g., mitotic activity) are no longer prognostic in *IDH* mutant astrocytomas (38). The field of neuro-oncology is currently awaiting features that can better discriminate more aggressive gliomas from those that are more indolent. Improving the accuracy and objectivity of grading will directly impact

Table 1. Hazard ratios for single- and multiple-variable Cox regression models

	Single variable				Multivariable (SCNN)			Multivariable (GSCNN)		
Variable	c Index	Hazard ratio	95% CI	*P* value	Hazard ratio	95% CI	*P* value	Hazard ratio	95% CI	*P* value
SCNN	**0.741**	**7.15**	**5.64, 9.07**	**2.08e-61**	**3.05**	**2.22, 4.19**	**2.71e-12**	—	—	—
GSCNN	0.781	12.60	9.34, 17.0	**3.08e-64**	—	—	—	8.83	4.66, 16.74	**9.69e-12**
IDH WT astrocytoma	0.726	9.21	6.88, 12.34	**3.48e-52**	4.73	2.57, 8.70	**3.49e-7**	0.97	0.43, 2.17	0.93
IDH mutant astrocytoma	—	0.23	0.170, 0.324	**2.70e-19**	2.35	1.27, 4.34	**5.36e-3**	1.67	0.90, 3.12	0.10
Histologic grade IV	0.721	7.25	5.58, 9.43	**2.68e-51**	1.52	0.839, 2.743	0.159	1.98	1.11, 3.51	**0.017**
Histologic grade III	—	0.44	0.332, 0.591	**1.66e-08**	1.57	0.934, 2.638	0.0820	1.78	1.07, 2.97	**0.024**
Age	0.744	1.77	1.63, 1.93	**2.52e-42**	1.33	1.20, 1.47	**9.57e-9**	1.34	1.22, 1.48	**9.30e-10**
Sex, female	0.552	0.89	0.706, 1.112	0.29	0.85	0.67, 1.08	0.168	0.86	0.68, 1.08	0.18

Bold indicates statistical significance ($P < 5e\text{-}2$).

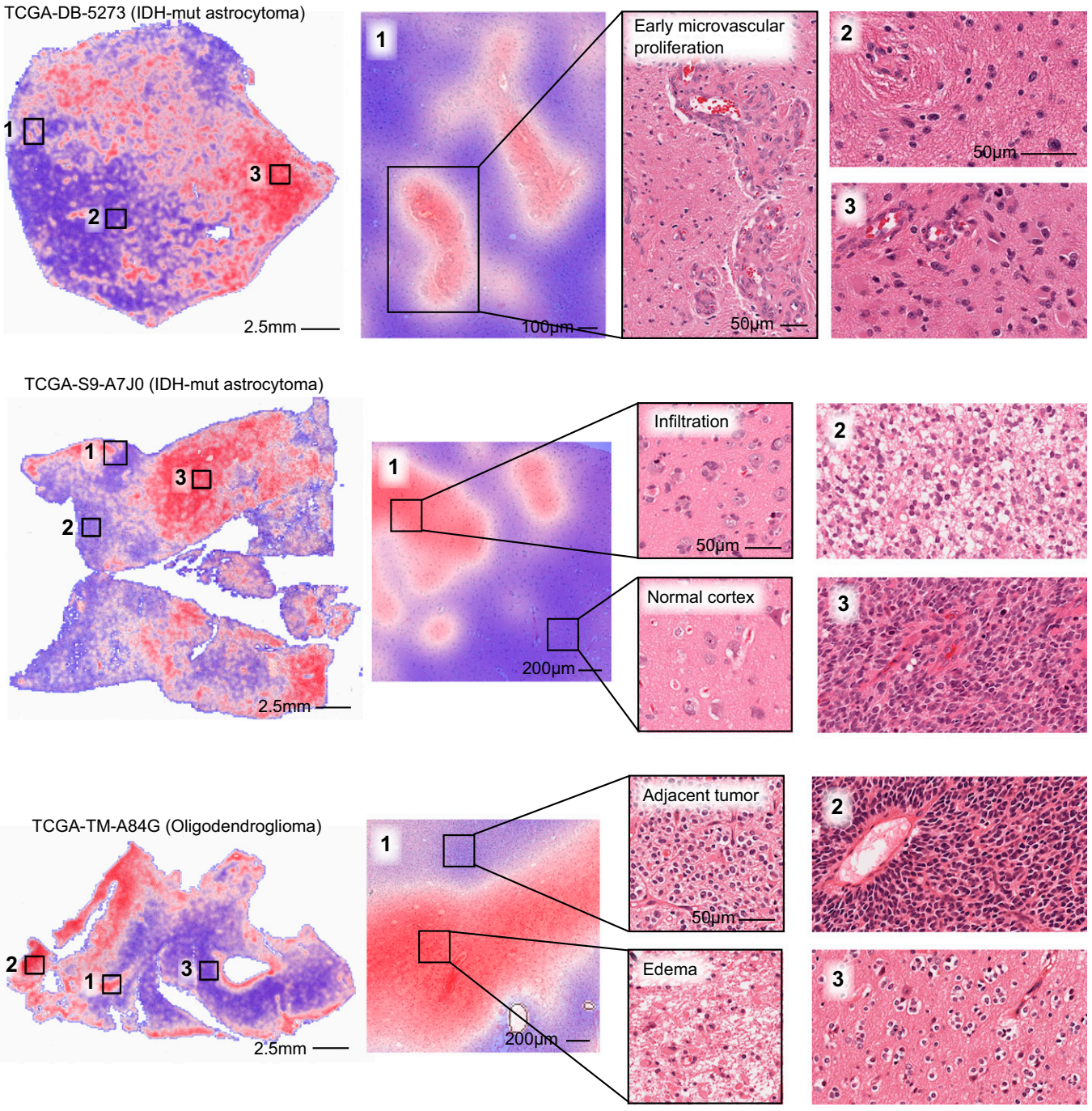

Fig. 5. Visualizing risk with whole-slide SCNN heat maps. We performed SCNN predictions exhaustively within whole-slide images to generate heat map overlays of the risks that SCNN associates with different histologic patterns. Red indicates relatively higher risk, and blue indicates lower risk (the scale for each slide is different). (*Top*) In TCGA-DB-5273, SCNN clearly and specifically predicts high risks for regions of early microvascular proliferation (region 1) and also, higher risks with increasing tumor infiltration and cell density (region 2 vs. 3). (*Middle*) In TCGA-S9-A7J0, SCNN can appropriately discriminate between normal cortex (region 1 in *Bottom*) and adjacent regions infiltrated by tumor (region 1 in *Top*). Highly cellular regions containing prominent microvascular structures (region 3) are again assigned higher risks than lower-density regions of tumor (region 2). Interestingly, low-density infiltrate in the cortex was associated with high risk (region 1 in *Top*). (*Bottom*) In TCGA-TM-A84G, SCNN assigns high risks to edematous regions (region 1 in *Bottom*) that are adjacent to tumor (region 1 in *Top*).

patient care by identifying patients who can benefit from more aggressive therapeutic regimens and by sparing those with less aggressive disease from unnecessary treatment.

Remarkably, SCNN performed as well as manual histologic grading or molecular subtyping in predicting overall survival in our dataset, despite using only a very small portion of each histology image for training and prediction. Additional investigation of the associations between SCNN risk predictions, molecular subtypes, and histologic grades revealed that SCNN can effectively discriminate outcomes within each molecular subtype, effectively performing digital histologic grading. Furthermore, SCNN can effectively recognize histologic differences associated with *IDH* mutations in astrocytomas and predict outcomes for these patients accordingly. SCNNs correctly predicted lower

risks for WHO grade III *IDH* mutant astrocytomas compared with WHO grade III *IDH* WT astrocytomas, consistent with the considerably longer median survival for patients with *IDH* mutant astrocytoma (6.3 vs. 1.7 y). While there are histologic features of astrocytomas that are understood to be more prevalent in *IDH* mutant astrocytomas, including the presence of microcysts and the rounded nuclear morphology of neoplastic nuclei, these are not reliable predictors of *IDH* mutations (40).

To integrate genomic information in prognostication, we developed a hybrid network that can learn simultaneously from both histology images and genomic biomarkers. The GSCNN presented in our study significantly outperforms the WHO standard based on identical inputs. We compared the performance of GSCNN and SCNN in several ways to evaluate their ability to predict survival and to assess the relative importance of histology and genomic data in GSCNN. GSCNN had significantly higher c index scores due to the inclusion of genomic variables in the training process. Performance significantly declined when using a superficial integration method that combines genomic biomarkers with a pretrained SCNN model.

In multivariable regression analyses, GSCNN has a much higher hazard ratio than SCNN (8.83 vs. 3.05). Examining the other variables in the regression models, we noticed an interesting relationship between the significance of histologic-grade and molecular subtype variables. In the SCNN regression analysis, histologic-grade variables were not significant, but molecular subtype variables were highly significant, indicating that SCNN could capture histologic information from image data but could not learn molecular subtype information entirely from histology. In contrast, molecular subtype information was not significant in the GSCNN regression analysis. Interestingly, histologic-grade variables were marginally significant, suggesting that some prognostic value in the histology images remained untapped by GSCNN.

Kaplan–Meier analysis showed remarkable similarity in the discriminative power of SCNN and GSCNN. Additional Kaplan–Meier analysis of risk categories within molecular subtypes revealed interesting trends that are consistent with the regression analyses presented in Table 1. SCNN clearly separates outcomes within each molecular subtype based on histology. Survival curves for GSCNN risk categories, however, overlap significantly in each subtype. Since SCNN models do not have access to genomic data when making predictions, their ability to discriminate outcomes was worse in general when assessed by c index or multivariable regression.

Integration of genomic and histology data into a single prediction framework remains a challenge in the clinical implementation of computational pathology. Our previous work in developing deep learning survival models from genomic data has shown that accurate survival predictions can be learned from high-dimensional genomic and protein expression signatures (29). Incorporating additional genomic variables into GSCNN models is an area for future research and requires larger datasets that combine histology images with rich genomic and clinical annotations.

While deep learning methods frequently deliver outstanding performance, the interpretability of black box deep learning models is limited and remains a significant barrier in their validation and adoption. Heat map analysis provides insights into the histologic patterns associated with increased risk and can also serve as a practical tool to guide pathologists to tissue regions associated with worse prognosis. The heat maps suggest that SCNN can learn visual patterns known to be associated with histologic features related to prognosis and used in grading, including microvascular proliferation, cell density, and nuclear morphology. Microvascular prominence and proliferation are associated with disease progression in all forms of diffuse glioma, and these features are clearly delineated as high risk in the heat map presented for slide TCGA-DB-5273. Likewise, increases in cell density and nuclear pleomorphism were also associated with increased risk in all examples. SCNN also assigned high risks to regions that do not contain well-recognized features associated with a higher grade or poor prognosis. In region 1 of slide TCGA-S9-A7J0, SCNN assigns higher risk to sparsely infiltrated cerebral cortex than to region 2, which is infiltrated by a higher density of tumor cells (normal cortex in region 1 is properly assigned a very low risk). Widespread infiltration into distant sites of the brain is a hallmark of gliomas and results in treatment failure, since surgical resection of visible tumor often leaves residual neoplastic infiltrates. Similarly, region 1 of slide TCGA-TM-A84G illustrates a high risk associated with low-cellularity edematous regions compared with adjacent oligodendroglioma with much higher cellularity. Edema is frequently observed within gliomas and in adjacent brain, and its degree may be related to the rate of growth (43), but its histologic presence has not been previously recognized as a feature of aggressive behavior or incorporated into grading paradigms. While it is not entirely clear why SCNN assigns higher risks to the regions in the sparsely infiltrated or edematous regions, these examples confirm that SCNN risks are not purely a function of cellular density or nuclear atypia. Our human interpretations of these findings provide possible explanations for why SCNN unexpectedly predicts high risks in these regions, but these findings need additional investigation to better understand what specific features the SCNN network perceives in these regions. Nevertheless, this shows that SCNN can be used to identify potentially practice-changing features associated with increased risk that are embedded within pathology images.

Although our study provides insights into the application of deep learning in precision medicine, it has some important limitations. A relatively small portion of each slide was used for training and prediction, and the selection of ROIs within each slide required expert guidance. Future studies will explore more advanced methods for automatic selection of regions and for incorporating a higher proportion of each slide in training and prediction to better account for intratumoral heterogeneity. We also plan to pursue the development of enhanced GSCNN models that incorporate additional molecular features and to evaluate the value added of histology in these more complex models. In our Kaplan–Meier analysis, the thresholds used to define risk categories were determined in a subjective manner using the proportion of manual histologic grades in the TCGA cohort, and a larger dataset would permit a more rigorous definition of these thresholds to optimize survival stratification. The interpretation of risk heat maps was based on subjective evaluation by neuropathologists, and we plan to pursue studies that evaluate heat maps in a more objective manner to discover and validate histologic features associated with poor outcomes. Finally, while we have applied our techniques to gliomas, validation of these approaches in other diseases is needed and could provide additional insights. In fact, our methods are not specific to histology imaging or cancer applications and could be adapted to other medical imaging modalities and biomedical applications.

Methods

Data and Image Curation. Whole-slide images and clinical and genomic data were obtained from TCGA via the Genomic Data Commons (https://gdc.cancer.gov/). Images of diagnostic H&E-stained, formalin-fixed, paraffin-embedded sections from the Brain LGG and the GBM cohorts were reviewed to remove images containing tissue-processing artifacts, including bubbles, section folds, pen markings, and poor staining. Representative ROIs containing primarily tumor nuclei were manually identified for each slide that passed a quality control review. This review identified whole-slide images with poor image quality arising from imaging artifacts or tissue processing (bubbles, significant tissue section folds, overstaining, understaining) where suitable ROIs could not be selected. In the case of grade IV disease, some regions include microvascular proliferation, as this feature was exhibited throughout tumor regions. Regions containing geographic necrosis

were excluded. A total of 1,061 whole-slide images from 769 unique patients were analyzed.

ROI images (1,024 × 1,024 pixels) were cropped at 20× objective magnification using OpenSlide and color-normalized to a gold standard H&E calibration image to improve consistency of color characteristics across slides. HPFs at 256 × 256 pixels were sampled from these regions and used for training and testing as described below.

Network Architecture and Training Procedures. The SCNN combines elements of the 19-layer Visual Geometry Group (VGG) convolutional network architecture with a Cox proportional hazards model to predict time-to-event data from images (Fig. S1) (44). Image feature extraction is achieved by four groups of convolutional layers. (*i*) The first group contains two convolutional layers with 64 3 × 3 kernels interleaved with local normalization layers and then followed with a single maximum pooling layer. (*ii*) The second group contains two convolutional layers (128 3 × 3 kernels) interleaved with two local normalization layers followed by a single maximum pooling layer. (*iii*) The third group interleaves four convolutional layers (256 3 × 3 kernels) with four local normalization layers followed by a single maximum pooling layer. (*iv*) The fourth group contains interleaves of eight convolutional (512 3 × 3 kernels) and eight local normalization layers, with an intermediate pooling layer and a terminal maximum pooling layer. These four groups are followed by a sequence of three fully connected layers containing 1,000, 1,000, and 256 nodes, respectively.

The terminal fully connected layer outputs a prediction of risk $R = \beta^T X$ associated with the input image, where $\beta \in \mathbb{R}^{256\times 1}$ are the terminal layer weights and $X \in \mathbb{R}^{256\times 1}$ are the inputs to this layer. To provide an error signal for backpropagation, these risks are input to a Cox proportional hazards layer to calculate the negative partial log likelihood:

$$L(\beta, X) = -\sum_{i\in U}\left(\beta^T X_i - \log\sum_{j\in\Omega_i} e^{\beta^T X_j}\right), \quad [1]$$

where $\beta^T X_i$ is the risk associated with HPF i, U is the set of right-censored samples, and Ω_i is the set of "at-risk" samples with event or follow-up times $\Omega_i = \{j | Y_j \geq Y_i\}$ (where Y_i is the event or last follow-up time of patient i).

The adagrad algorithm was used to minimize the negative partial log likelihood via backpropagation to optimize model weights, biases, and convolutional kernels (45). Parameters to adagrad include the initial accumulator value = 0.1, initial learning rate = 0.001, and an exponential learning rate decay factor = 0.1. Model weights were initialized using the variance scaling method (46), and a weight decay was applied to the fully connected layers during training (decay rate = 4e-4). Models were trained for 100 epochs (1 epoch is one complete cycle through all training samples) using minibatches consisting of 14 HPFs each. Each minibatch produces a model update, resulting in multiple updates per epoch. Calculation of the Cox partial likelihood requires access to the predicted risks of all samples, which are not available within any single minibatch, and therefore, Cox likelihood was calculated locally within each minibatch to perform updates (U and Ω_i were restricted to samples within each minibatch). Local likelihood calculation can be very sensitive to how samples are assigned to minibatches, and therefore, we randomize the minibatch sample assignments at the beginning of each epoch to improve robustness. Mild regularization was applied during training by randomly dropping out 5% of weights in the last fully connected layer in each minibatch during training to mitigate overfitting.

Training Sampling. Each patient has possibly multiple slides and multiple regions within each slide that can be used to sample HPFs. During training, a single HPF was sampled from each region, and these HPFs were treated as semiindependent training samples. Each HPF was paired with patient outcome for training, duplicating outcomes for patients containing multiple regions/HPFs. The HPFs are sampled at the beginning of each training epoch to generate an entirely new set of HPFs. Randomized transforms were also applied to these HPFs to improve robustness to tissue orientation and color variations. Since the visual patterns in tissues can often be anisotropic, we randomly apply a mirror transform to each HPF. We also generate random transformations of contrast and brightness using the "random_contrast" and "random_brightness" TensorFlow operations. The contrast factor was randomly selected in the interval [0.2, 1.8], and the brightness was randomly selected in the interval [−63, 63]. These sampling and transformation procedures along with the use of multiple HPFs for each patient have the effect of augmenting the effective size of the labeled training data. In tissues with pronounced anisotropy, including adenocarcinomas that exhibit prominent glandular structures, these mirror transformations are intended to improve the robustness of the network to tissue orientation. Similar approaches for training data augmentation have shown considerable improvements in general imaging applications (33).

Testing Sampling, Risk Filtering, and Model Averaging. Sampling was also performed to increase the robustness and stability of predictions. (*i*) Nine HPFs are first sampled from each region j corresponding to patient m. (*ii*) The risk of the kth HPF in region j for patient m, denoted $R_m^{j,k}$, is then calculated using the trained SCNN model. (*iii*) The median risk $R_m^j = median_k\{R_m^{j,k}\}$ is calculated for region j using the aforementioned HPFs to reject outlying risks. (*iv*) These median risks are then sorted from highest to lowest $\widehat{R_m^1} > \widehat{R_m^2} > \widehat{R_m^3} \ldots$, where the superscript index now corresponds to the risk rank. (*v*) The risk prediction for patient m is then selected as the second highest risk $R_m^* = \widehat{R_m^2}$. This filtering procedure was designed to emulate how a pathologist integrates information from multiple areas within a slide, determining prognosis based on the region associated with the worst prognosis. Selection of the second highest risk (as opposed to the highest risk) introduces robustness to outliers or high risks that may occur due to some imaging or tissue-processing artifact.

Since the accuracy of our models can vary significantly from one epoch to another, largely due to the training sampling and randomized minibatch assignments, a model-averaging technique was used to reduce prediction variance. To obtain final risk predictions for the testing patients that are stable, we perform model averaging using the models from epochs 96 to 100 to smooth variations across epochs and increase stability. Formally, the model-averaged risk for patient m is calculated as

$$\overline{R_m^*} = \frac{1}{5}\sum_{\gamma=96}^{100} R_{m(\gamma)}^*, \quad [2]$$

where $R_{m(\gamma)}^*$ denotes the predicted risk for patient m in training epoch γ.

Validation Procedures. Patients were randomly assigned to nonoverlapping training (80%) and test (20%) sets that were used to train models and evaluate their performance. If a patient was assigned to training, then all slides corresponding to that patient were assigned to the training set and likewise, for the testing set. This ensures that no data from any one patient are represented in both training and testing sets to avoid overfitting and optimistic estimates of generalization accuracy. We repeated the randomized assignment of patients training/testing sets 15 times and used each of these training/testing sets to train and evaluate a model. The same training/testing assignments were used in each model (SCNN, GSCNN, baseline) for comparability. Prediction accuracy was measured using Harrell's c index to measure the concordance between predicted risk and actual survival for testing samples (36).

Statistical Analyses. The c indices generated by Monte Carlo cross-validation were performed using the Wilcoxon signed rank test. This paired test was chosen, because each method was evaluated using identical training/testing sets. Comparisons of SCNN risk values across grade were performed using the Wilcoxon rank sum test. Cox univariable and multivariable regression analyses were performed using predicted SCNN risk values for all training and testing samples in randomized training/testing set 1. Analyses of the correlation of grade, molecular subtype, and SCNN risk predictions were performed by pooling predicted risks for testing samples across all experiments. SCNN risks were normalized within each experiment by z score before pooling. Grade analysis was performed by determining "digital"-grade thresholds for SCNN risks in each subtype. Thresholds were objectively selected to match the proportions of samples in each histologic grade in each subtype. Statistical analysis of Kaplan–Meier plots was performed using the log rank test.

Hardware and Software. Prediction models were trained using TensorFlow (v0.12.0) on servers equipped with dual Intel(R) Xeon(R) CPU E5-2630L v2 @ 2.40 GHz CPUs, 128 GB RAM, and dual NVIDIA K80 graphics cards. Image data were extracted from Aperio .svs whole-slide image formats using OpenSlide (openslide.org/). Basic image analysis operations were performed using HistomicsTK (https://github.com/DigitalSlideArchive/HistomicsTK), a Python package for histology image analysis.

Data Availability. This paper was produced using large volumes of publicly available genomic and imaging data. The authors have made every effort to make available links to these resources as well as make publicly available the software methods and information used to produce the datasets, analyses, and summary information.

ACKNOWLEDGMENTS. This work was supported by US NIH National Library of Medicine Career Development Award K22LM011576 and National Cancer Institute Grant U24CA194362 and by the National Brain Tumor Society.

1. Kong J, et al. (2008) Computer-assisted grading of neuroblastic differentiation. *Arch Pathol Lab Med* 132:903–904, author reply 904.
2. Niazi MKK, et al. (2017) Visually meaningful histopathological features for automatic grading of prostate cancer. *IEEE J Biomed Health Inform* 21:1027–1038.
3. Naik S, et al. (2008) Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. *Proceedings of the 2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro* (IEEE, Piscataway, NJ), pp 284–287.
4. Ren J, et al. (2015) Computer aided analysis of prostate histopathology images Gleason grading especially for Gleason score 7. *Conf Proc IEEE Eng Med Biol Soc* 2015: 3013–3016.
5. Kothari S, Phan JH, Young AN, Wang MD (2013) Histological image classification using biologically interpretable shape-based features. *BMC Med Imaging* 13:9.
6. Sertel O, et al. (2009) Computer-aided prognosis of neuroblastoma on whole-slide images: Classification of stromal development. *Pattern Recognit* 42:1093–1103.
7. Fauzi MF, et al. (2015) Classification of follicular lymphoma: the effect of computer aid on pathologists grading. *BMC Med Inform Decis Mak* 15:115.
8. Dundar MM, et al. (2011) Computerized classification of intraductal breast lesions using histopathological images. *IEEE Trans Biomed Eng* 58:1977–1984.
9. Hou L, et al. (2016) Patch-based convolutional neural network for whole slide tissue image classification. *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Piscataway, NJ), pp 2424–2433.
10. Kong J, et al. (2013) Machine-based morphologic analysis of glioblastoma using whole-slide pathology images uncovers clinically relevant molecular correlates. *PLoS One* 8:e81049.
11. Wang D, Khosla A, Gargeya R, Irshad H, Beck AH (2016) Deep learning for identifying metastatic breast cancer. arXiv:1606.05718.
12. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521:436–444.
13. Greenspan H, van Ginneken B, Summers RM (2016) Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Trans Med Imaging* 35:1153–1159.
14. Janowczyk A, Madabhushi A (2016) Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *J Pathol Inform* 7:29.
15. Litjens G, et al. (2016) Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci Rep* 6:26286.
16. Chen T, Chefd'hotel C (2014) Deep learning based automatic immune cell detection for immunohistochemistry images. *Machine Learning in Medical Imaging* (Springer, Berlin), pp 17–24.
17. Cruz-Roa A, et al. (2017) Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. *Sci Rep* 7:46450.
18. Pereira S, Pinto A, Alves V, Silva CA (2016) Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Trans Med Imaging* 35:1240–1251.
19. Sirinukunwattana K, et al. (2016) Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Trans Med Imaging* 35:1196–1206.
20. Esteva A, et al. (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542:115–118.
21. Gulshan V, et al. (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 316: 2402–2410.
22. Havaei M, et al. (2017) Brain tumor segmentation with deep neural networks. *Med Image Anal* 35:18–31.
23. Huynh BQ, Li H, Giger ML (2016) Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *J Med Imaging (Bellingham)* 3:034501.
24. Kamnitsas K, et al. (2017) Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal* 36:61–78.
25. Turkki R, Linder N, Kovanen PE, Pellinen T, Lundin J (2016) Antibody-supervised deep learning for quantification of tumor-infiltrating immune cells in hematoxylin and eosin stained breast cancer samples. *J Pathol Inform* 7:38.
26. Bychkov D, Turkki R, Haglund C, Linder N, Lundin J (2016) Deep learning for tissue microarray image-based outcome prediction in patients with colorectal cancer. *SPIE Medical Imaging*, eds Gurcan MN, Madabhushi A (International Society for Optics and Photonics, Bellingham, WA), p 6.
27. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI (2014) Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J* 13:8–17.
28. Xiang A, Lapuerta P, Ryutov A, Buckley J, Azen S (2000) Comparison of the performance of neural network methods and Cox regression for censored survival data. *Comput Stat Data Anal* 34:243–257.
29. Yousefi S, et al. (2017) Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Sci Rep* 7:11707.
30. Yousefi S, Congzheng S, Nelson N, Cooper LAD (2016) Learning genomic representations to predict clinical outcomes in cancer. arXiv:1609.08663.
31. Katzman J, et al. (2016) DeepSurv: Personalized treatment recommender system using A Cox proportional hazards deep neural network. arXiv:1606.00931.
32. Zhu X, Yao J, Huang J (2016) Deep convolutional neural network for survival analysis with pathological images. *Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine* (IEEE, Piscataway, NJ), pp 544–547.
33. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, eds Pereira F, Burges CJC, Bottou L, Weinberger KQ (Neural Information Processing Systems Foundation, Inc., La Jolla, CA), pp 1097–1105.
34. Gutman DA, et al. (2013) Cancer Digital Slide Archive: an informatics resource to support integrated in silico analysis of TCGA pathology data. *J Am Med Inform Assoc* 20:1091–1098.
35. Gutman DA, et al. (2017) The digital slide archive: A software platform for management, integration, and analysis of histology for cancer research. *Cancer Res* 77: e75–e78.
36. Harrell FE, Jr, Califf RM, Pryor DB, Lee KL, Rosati RA (1982) Evaluating the yield of medical tests. *JAMA* 247:2543–2546.
37. Brat DJ, et al.; Cancer Genome Atlas Research Network (2015) Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N Engl J Med* 372: 2481–2498.
38. Reuss DE, et al. (2015) IDH mutant diffuse and anaplastic astrocytomas have similar age at presentation and little difference in survival: a grading problem for WHO. *Acta Neuropathol* 129:867–873.
39. Leeper HE, et al. (2015) IDH mutation, 1p19q codeletion and ATRX loss in WHO grade II gliomas. *Oncotarget* 6:30295–30305.
40. Nguyen DN, et al. (2013) Molecular and morphologic correlates of the alternative lengthening of telomeres phenotype in high-grade astrocytomas. *Brain Pathol* 23: 237–243.
41. Wijnenga MMJ, et al. (2018) The impact of surgery in molecularly defined low-grade glioma: an integrated clinical, radiological, and molecular analysis. *Neuro-oncol* 20: 103–112.
42. van den Bent MJ (2010) Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician's perspective. *Acta Neuropathol* 120:297–304.
43. Pope WB, et al. (2005) MR imaging correlates of survival in patients with high-grade gliomas. *AJNR Am J Neuroradiol* 26:2466–2474.
44. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.
45. Duchi J, Hazan E, Singer Y (2011) Adaptive subgradient methods for online learning and stochastic optimization. *J Mach Learn Res* 12:2121–2159.
46. He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *IEEE International Conference on Computer Vision* (IEEE, Piscataway, NJ), pp 1026–1034.