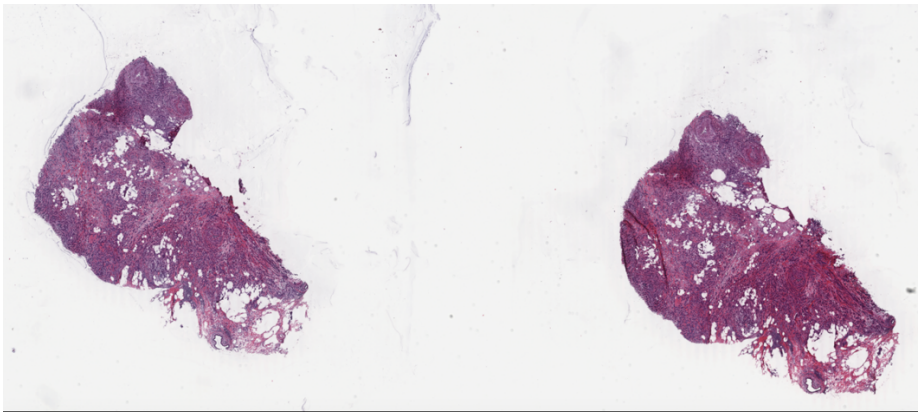
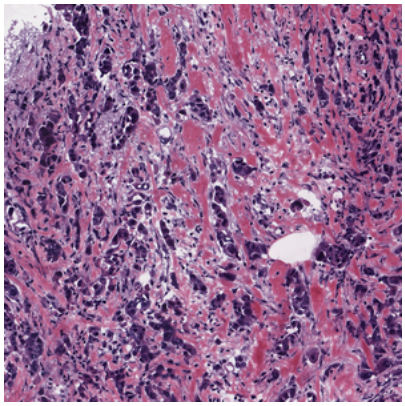
داده‌ی مورد نیاز برای پروژه از دو قسمت متن گزارش و تصاویر مرتبط با آن گزارش، تشکیل شده‌است. برای استفاده و دانلود تصاویر نیاز است UUID هر کدام از تصاویر را پیدا کنیم. برای این کار manifest کامل دیتاست مورد نظر را دانلود می‌کنیم که شامل UUID، نام فایل، سایز فایل و موارد دیگر است. از روی نام فایل آن‌هایی که فرمت svs دارند و جزو گزارش‌های انتخاب‌شده هستند را فیلتر می‌کنیم. نهایتا یک فایل manifest خواهیم داشت که تمام موارد آن باید دانلود و پردازش شود.

با توجه به این که تصاویر موردنظر به صورت مستقیم از منبع انتخاب شده قابلیت دانلود شدن را ندارند، نیاز است برنامه‌ی gdc client نصب شود. این برنامه در دو نسخه‌ی دارای محیط گرافیکی و بدون محیط گرافیکی ارائه می‌شود که با توجه به فرآیند انجام کار که در colab اجرا می‌شود، نسخه‌ی بدون محیط گرافیکی نصب شده‌است. این برنامه UUID را به عنوان ورودی گرفته و فایل مورد نظر را دانلود و ذخیره می‌کند. با توجه به فایل manifest ساخته شده، فایل‌ها را یکی یکی دانلود می‌کنیم. مجموع حجم فایل‌های دانلود شده حدودا ۶۰ گیگابایت است.



نمونه‌ای از تصویر به صورت کامل



نمونه‌ای از پچ‌های ساخته شده

برای خواندن تصاویر با فرمت svs، از کتابخانه‌ی openslide استفاده کرده‌ایم. این کتابخانه تابعی را معرفی می‌کند که با فراخوانی آن می‌توانیم بخشی از تصویر را در پایتون برای پردازش کردن، بارگذاری کنیم. امکان بارگذاری کامل تصویر با توجه به حجم زیاد آن و محدودیت حافظه‌ی colab، وجود ندارد.

با توجه به هدف پروژه، نیاز است تصاویر دانلود شده به صورت پچ‌هایی با سایز کمتر تبدیل شوند. برای این کار روی عکس در جهت سطری و در جهت ستونی با گام‌های ۱۰۲۴ پیکسلی حرکت می‌کنیم و ناحیه انتخاب

شده را با استفاده از openslide می‌خوانیم. با این کار عکس مورد نظر به پچ‌هایی با سایز ۱۰۲۴ در ۱۰۲۴ پیکسل تبدیل می‌شود. از آنجا که همه‌ی این پچ‌ها حاوی شی مورد نظر نیستند و اطلاعات مهمی در خود ندارند، نیاز است از مجموعه‌ی پچ‌های انتخاب شده حذف شوند. برای این‌کار از الگوریتم آماده‌ی otsu استفاده می‌کنیم. این الگوریتم تصویر را به دو دسته‌ی پس‌زمینه و پیش‌زمینه، تقسیم می‌کند. اگر درصد زیادی از تصویر مورد نظر به پس زمینه، که سفید است، تعلق داشته باشد، آن را حذف می‌کنیم. در نهایت پچ‌هایی باقی خواهد ماند که حاوی اطلاعات مهم هستند. این کار را بار دیگر با گام ۲۰۴۸ و ۳۰۷۲ پیکسل انجام می‌دهیم.